Approaches to systems biology. Four methods to study single-cell gene expression, cell motility, antibody reactivity, and respiratory metabolism

Hagedorn, Peter

Publication date:
2006

Document Version
Publisher's PDF, also known as Version of record

Risø-PhD-23(EN)

Approaches to systems biology

Four methods to study single-cell gene expression, cell motility, antibody reactivity, and respiratory metabolism

Peter H. Hagedorn

Risø National Laboratory
Roskilde
Denmark
March 2006

Ph.D. thesis statement:
This thesis is submitted in partial fulfilment of the requirements for the Ph.D. degree at the Department of Biochemistry and Molecular Biology, Faculty of Science and Engineering, University of Southern Denmark.

Author:
Peter H. Hagedorn
Biosystems Department
Risø National Laboratory

Title:
Approaches to systems biology
— Four methods to study single-cell gene expression, cell motility, antibody reactivity, and respiratory metabolism

Abstract: See pages i and ii.

Supervisors:
Henrik Flyvbjerg
Senior scientist, dr. scient
Biosystems Department
Risø National Laboratory

Ian Max Møller
Professor
Department of Agricultural Sciences
The Royal Veterinary and Agricultural University

Michael Foged Lyngkjær,
Senior scientist
Biosystems Department
Risø National Laboratory

Lars Folke Olsen,
Associate professor
Department of Biochemistry and Molecular Biology
University of Southern Denmark

Report number:
Risø-PhD-23(EN)

ISBN: 87-550-3503-5

Reg. no.: 10039-3

Pages: 175
Tables: 2
Figures: 21
References: 157

Risø National Laboratory
Information Service Department
P.O.Box 49
DK-4000 Roskilde
Denmark
Telephone +45 4677 4004
bibl@risoe.dk
Fax +45 4677 4013
www.risoe.dk

Abstract

To understand how complex systems, such as cells, function, comprehensive measurements of their constituent parts must be made. This can be achieved by combining methods that are each optimized to measure specific parts of the system. Four such methods, each covering a different area, are presented: Transcript profiling of one cell type extracted from a complex tissue containing several cell types; observation and recording of cell motility; measurement of antibody reactivities using microarrays; and in vivo measurement of free and bound NADH in mitochondria.

Detailed statistical analysis of the data from such measurements allows models of the system to be developed and tested. For each of the methods, such analysis and modelling approaches have been applied and are presented: Differentially regulated genes are identified and classified according to function; cell-specific motility models are developed that can distinguish between different surfaces; a method for selecting repertoires of antigens that separate mice based on their response to treatment is developed; and the observed concentrations of free and bound NADH is used to build and test a basic model of respiratory metabolism.

The significance of each of the methods is exemplified by specific results and insights into the systems analyzed. But also, it is discussed how the methods relate to high-throughput biology in general, and how they (and similar methods) may be combined to rationalize the structure and kinetics of complex networks of interacting parts in terms of function.

Abstract in danish

For at forstå hvordan komplekse systemer, f.eks. celler, fungerer, må der udføres omfattende målinger af alle de enkelte dele i systemet. Dette kan f.eks. gøres ved at kombinere metoder, der hver især er optimeret til at måle specifikke dele af systemet. Fire sådanne metoder, der hver især dækker ét specifikt område, præsenteres her: Transkript profilering af en celle type, ekstraheret fra et komplekst væv der består af flere forskellige celle typer; observation af celle motilitet; måling af antistof reaktiviteter vha. microarrays; og in vivo målinger af frit og bundet NADH i mitokondrier.

Detaljeret statistisk analyse af data fra sådanne målinger gør det muligt at udvikle og teste modeller for systemerne. For hver af de fire metoder er en sådan analyse og modellering blevet anvendt og præsenteres her: Differentielt regulerede gener identificeres og klassificeres efter funktion; celle specifikke motilitets modeller, der kan se forskel på forskellige overflader, bliver udviklet; en metode til at udvælge grupper af antigener der kan adskille mus på baggrund af deres behandlings respons udvikles; og de observerede koncentrationer af frit og bundet NADH bliver brugt til at udvikle og teste en simpel model for den respiratoriske metabolisme.

Betydningen af hver af de udviklede metoder eksemplificeres ved de specifikke resultater og indsigter, der er opnået for hver af de systemer, der bliver analyseret. Derudover diskuteres det, hvorledes de præsenterede metoder generelt relaterer til moderne biologi, hvor mange parametre måles samtidigt, og hvorledes de (og andre metoder) kan kombineres for at rationalisere udformningen og kinetikken i komplekse netværk af interagerende dele ud fra funktion.

Acknowledgements

It is a pleasure to thank my supervisors for their guidance, patience, and for the amount of time they have invested in my education and me. Henrik Flyvbjerg, for introducing me to science and for trying to teach me how to navigate in the world of science. Ian Max Møller, for providing me with an understanding of the importance of the "bio" in biophysics, and helping me develop models that incorporate this, and for your particular interest in improving my presentation and writing skills. Michael Lyngkjær, for letting me be a member of your group, for showing me how fun it can be to work with gene expression, and for the amount of confidence you have always had in my ideas. Lars Folke Olsen, for always being so pleasant and forthcoming when details about administrative aspects of the Ph.D. needed to be discussed, and for providing such good advice on the structuring of the thesis. Also, Eytan Domany and Irun Cohen, for making my stay at the Weizmann Institute, Israel, so rewarding and interesting. My keen interest in bioinformatics largely stems from working with you.

For being so pleasant and productive to work with: Francisco Quintana, Torben Gjetting, Marina Kasimova, Klaas Krab, Gad Elizur and David Selmeczi. For providing a stable work-environment: Hans Thordal-Christensen and Iver Jakobsen. For the insights gained from supervising your projects: Jens. V. Johansen, Peter Hjorth, Kaare Graesboell, Andreas Blicher, Tommy Rex Gravesen, and Sofie Inari Castella. For the many discussions on the meaning of it all: Andrew C. Mumm — I sincerely thank you all!

I would also like to thank my family, in particular my wife Marie. I am so grateful you have put up with the stresses and strains of the past three years. Without your gentle and loving encouragement I would never have made it. Also, my son Marius, for putting everything so clearly in perspective. Only the joys you can share with the ones you love matter.

Peter H. Hagedorn
Copenhagen
March, 2006

Preface

After three years at Risø National Laboratory (RISØ), one principle has been constantly emerging in all the projects I have participated in: How fruitful the interplay between experimental and theoretical work can be. Only thorough analysis and theoretical treatment of experimental results give the insights needed to direct further improvements and optimization of the experimental procedures. And only comprehensive understanding of the experiment and underlying biology allows interesting models to be build, tested, and adjusted. In the development phase, several cycles of experiments, theory, adjusted experiments, adjusted theory, etc., are needed before substantial results are achieved. When stated, this observation might seem trivial. However, to me, being most inclined to work on the theoretical and analytical parts of projects, it represent an insight I believe will increase the quality of my future work.

The work environment and supervision I have received at RISØ has encouraged me to work on several projects with different groups, mostly biologists. Consequently, I have not had one large project, but instead several smaller ones; some finished and published now, some still in progress, and some having produced nothing new. This has been very educational, scientifically as well as towards developing social interaction skills in multidisciplinary environments. With respect to presenting a coherent thesis, however, it has been a bit of a challenge. Still, several common themes are present in all of the projects, as I expand upon in the introduction, and I have structured the thesis around these themes.

Four projects are presented in the thesis, the published items of which are presented in the next section (Publications), and referred to throughout the text as papers I, II, III, and IV. For some of the papers, more than one item has been published or filed, and these are also listed. However, I use the term "paper", instead of "project", since there is only one peer-reviewed publication per project. The peer-reviewed publications are included at the end of the thesis (Enclosed publications).

When choosing topics and the level of detail to present them at, I have been guided by what I would myself have benefitted the most from reading three years ago when I started the research — with the additional constraint that the thesis should be kept at a reasonable size.

Publications

I. *Reviewed article*
Gjetting, T.*, P. H. Hagedorn*, P. Schweizer, H. Thordal-Christensen, M. F. Lyngkjær (2006) Single-cell transcript profiling of barley attacked by the powdery mildew fungus (Submitted and under review in Plant Journal)
*shared first-authorship

II. *Reviewed article*
Selmeczi, D., S. Mosler, P. H. Hagedorn, N. B. Larsen, and H. Flyvbjerg (2005). Cell motility as persistent random motion: theories from experiments. *Biophys. J. 89(2)*, 912–931.

Book chapter
Selmeczi D., S. Tólic-Nørrelykke, E. Schäffer, P. H. Hagedorn, S. Mosler, K. Berg-Sørensen, N. B. Larsen, H. Flyvbjerg (2005) Brownian Motion after Einstein: Some new applications and new experiments. In Controlled Nanoscale Motion in Biological and Artificial Systems, Linke H, Månsson A (editors). Springer, in press.

Popular article
Flyvbjerg H., D. Selmeczi, P. H. Hagedorn, N. B. Larsen (2005) Brownske bevægelser og mikroorganismers motilitet. *Kvant 16(2)*, 17–20.

III. *Reviewed article*
Quintana, F. J., P. H. Hagedorn, G. Elizur, Y. Merbl, E. Domany, and I. R.Cohen (2004). Functional immunomics: Microarray analysis of IgG autoantibody repertoires predicts the future response of mice to induced diabetes. *Proc. Natl. Acad. Sci. USA 101*, 14615–14621.

Patent
Cohen I. R., Quintana F. J., E. Domany, G. Elizur, P. H. Hagedorn (2005) Antigen array and diagnostic uses thereof. US patent application, Application number: 20050260770, Serial no: 094142, Series code: 11, Filed: March 31, 2005.

IV. *Reviewed article*
Kasimova. M. R., J. Grigiene , K. Krab, P. H. Hagedorn, H. Flyvbjerg, P. E. Andersen, I. M. Møller (2006) The Free NADH Concentration Is Kept Constant in Plant Mitochondria under Different Metabolic Conditions. *Plant Cell 18*, 688-798.

Proceedings
Kasimova. M. R., K. Krab, J. Grigiene , P. E. Andersen, P. H. Hagedorn, H. Flyvbjerg, I. M. Møller (2004) Quantitative distinction between bound and free NADH in biological systems (2004) In Biomedical Vibrational Spectroscopy and Biohazard Detection Technologies, Mahadevan-Jansen A., M. G. Sowa, G. J. Puppels, Z. Gryczynski, T. Vo-Dinh, J. R. Lakowicz (editors). *Proceedings of the SPIE 5322*, 42–51.

To my wife and son
I never knew how much I missed you until you arrived

Contents

Introduction and outline of thesis work

The success of the natural sciences in explaining the world around us, and in sometimes even allowing us to manipulate it to our benefit, is based on a certain way of generating new knowledge: The reproducible observation, or measurement, of some aspect of nature, followed by an interpretation of the measurements — often within the framework of some theory that the measurements can support or reject. In some cases, one can approach this very stringently by building a mathematical model that can reproduce the essential trends in the measurements. The art of building models and matching these to the observed data is a central theme in the research for this thesis.

Another central theme is the development and application of experimental methods to observe how molecules interact with each other and their surroundings at the cellular level. The cell, first observed by Robert Hooke in 1665 using a compound microscope, has since been recognized as the basic building block of all living organisms: It stores and processes the hereditary information that defines the organism, and, through replication and differentiation, it performs all the specialized functions necessary to sustain the organism. Enclosed by a plasma membrane, the inside of the cell is like a biochemical factory, with specialized organelles performing a variety of functions, supported by a myriad of different molecules. These molecules interact with each other and their surroundings in complex networks, and consequently, in order to understand how the cell functions and responds to stimuli at a whole-system level, ideally one needs to be able to measure the amount of each and every molecule in the cell at the relevant points in time.

System-wide measurements strike at a third central theme of the research for this thesis: The development and use of computational methods to handle and analyze large amounts of data in an automated fashion. Large amounts of data have for a long time been the norm in research fields such as nuclear physics (for example, once it is operational around 2007, CERNs new particle accelerator, the Large Hadron Collider, is expected to produce around 15 million gigabytes of data per year) or meteorology (spacecrafts constantly scan the earth, creating hundreds of gigabytes of data daily), but have only recently appeared in biology with the advent of new experimental techniques like genome sequencing, microarrays, and real-time monitoring and tracking, to name a few.

The thesis is structured as follows: In chapter 1, methods for isolating and ob-

serving whole cells are reviewed. The cell's "phenotypic" response to external stimuli ranges from the simple observation of, for example successful penetration of the cell wall by fungi (paper I), to the more complex observation of motility patterns (paper II). Preparation of cell contents for measurement, specifically the amplification of DNA, is also treated. In chapter 2, experimental techniques available to measure the key molecules in the cell or in other well-defined systems are reviewed. This is divided into measurements at the genome and transcriptome (paper I), proteome (paper III), and metabolome (paper IV) levels. Having discussed methods for measuring, and thus generating data, in chapters 1 and 2, chapter 3 discusses methods for analyzing the data, specifically the subject of fitting models to data. Often, the results from such analyses are still complex and difficult directly to interpret biologically. Chapter 4 discusses various efforts to relate results to biological knowledge. With the advent of high-throughput methods, the need for structuring the body of biological knowledge in ways that are easily accessible by automated queries is paramount. Some of the initiatives that have been applied with success in the research for this thesis are presented. Each chapter couples general discussions about methods and techniques with research examples showing the use of the methods in the research done for this thesis.

With the three central themes: building and fitting models to data, developing methods to study cellular functions at a system-wide level, and generating large amounts of data for which efficient computational methods are needed, the research for this thesis in general lies within the field of systems biology research. Each of the papers for this thesis can be seen as an independent attempt to approach the true system-wide measurement of all relevant aspects of the cell, for which they need to be combined. Consequently, the title of the thesis work was coined: Approaches to systems biology.

Building on this, the thesis is concluded with a discussion of the implications and future directions for the research presented. For some of the projects, additional datasets have already been generated and are currently being analyzed. In the spirit of systems biology, attempts at integrating the various sources of data are presented.

The biological systems and problems studied in this thesis come from a variety of areas: Grain response to fungus attack, immune response to chemicals with respect to diabetes development, cell motility response to different biomaterials to detect biocompatibility, and mitochondrial response to addition of substrate with respect to the key metabolite NADH. This wide range of problems demonstrates the general applicability of the methods presented in the thesis. To help put the research examples presented in the following chapters in perspective, each of the projects is introduced below.

Single-cell gene expression (paper I)

High-throughput gene expression (or transcript profiling) is well established in both plant and animal systems (Holloway et al., 2002). However, a limiting factor in these experiments is the amount of material needed. To get enough material, large tissue samples are often used. These samples usually contain a multitude of different cell types in varying amounts, and consequently, only

transcript profiles that represent averages of the transcript profiles in the individual cell types are measured. Also, even if homogeneous tissue samples can be obtained, the individual cells might be in very different states due to, e.g., a differential response to external factors.

We have consequently developed a system where (1) individual cells of a given type are first identified by microscopy. (2) The contents of these cells are then extracted using micro-capillaries. And (3), the mRNA is isolated, labelled, and amplified to give amounts sufficient for subsequent detection when hybridizing to cDNA arrays.

We have tested our method using the interaction between barley and powdery mildew as the model system. Here, the epidermal cells on the barley leaf are of primary interest because the powdery mildew attempts to penetrate these cells to establish a haustoria (fungal feeding structure). Some epidermal cells resist the penetration by successfully forming a papilla (callose deposition in the cell wall) at the site of attempted penetration, and others are not successful and get infected by a haustoria. In addition to this, not all epidermal cells will have powdery mildew spores interacting with them.

In order to maximize the amount of useful knowledge gained by applying our method to this system, we have developed a cDNA array that is specifically suited to measuring transcript activity resulting from this interaction.

In this manner, by micro-extraction of cell-specific mRNA and subsequent cDNA array analysis, we have successfully obtained separate gene expression profiles for specific mildew-resistant and -infected barley cells. With the protocols we have developed, this can be done with as few as 20 cells. Thus, it has been possible to identify genes that are specifically regulated in infected cells and presumably involved in fungal establishment.

Cell motility (paper II)

Cell migration is essential in many physiological and pathological processes, e.g., embryogenesis, wound healing, inflammation, and metastasis. It is also essential to emerging medical technologies that rely on colonization of biomaterials by migrating cells. Many cellular signaling pathways that regulate migration have been described in recent years (Ridley et al., 2003). The mechano-chemistry of migration is also studied and modeled in subtle detail (Gaudet et al., 2003, Ponti et al., 2004, Lauffenburger and Horwitz, 1996, Joanny et al., 2003). Less is known about the migratory pattern that results from a cell's processing of all external stimuli. Although motility models for bacteria have evolved to a high degree of sophistication (Berg and Brown, 1972, Berg, 2000, 2003, Gerbal et al., 2000), phenomenological mathematical models for cells from higher organisms have, with few exceptions (Shenderow and Sheetz, 1997), remained simple.

Using experimental time series for trajectories of two types of motile cells (NHDF; normal human dermal fibroblast cells, and HaCaT; human keratinocyte cells) on several types of surfaces (glass, polystyrene, molecular collagen, and fibrillar collagen), we have developed cell-type-specific motility models. The two models reflect the cells' different roles in the organism, it seems, and show that a cell has a memory of past velocities. Also, differences in the motility pattern of a given cell type on different surfaces are reflected in different

values for model parameters. Consequently, these parameter values are quantitative cell-and-surface compatibility measures. When designing biomaterials, such compatibility measures may prove helpful.

Antibody reactivity (paper III)

The antibodies present in one's body reflects the history of one's past immunological experience. However, can the present autoantibody repertoire be consulted to predict resistance or susceptibility to the future development of an autoimmune disease? To answer this we developed an antigen microarray and used bioinformatic analysis to study a model of type 1 diabetes developing in non-obese diabetic male mice in which the disease was accelerated and synchronized by exposing the mice to cyclophosphamide at 4 weeks of age.

We obtained sera from 19 individual mice, treated the mice to induce cyclophosphamide accelerated diabetes (CAD), and found, as expected, that 9 mice became severely diabetic, whereas 10 mice permanently resisted diabetes. We again obtained serum from each mouse after CAD induction. We then analyzed the patterns of antibody reactivities in individual mice to 266 different antigens spotted on the microarray. A selected panel of 27 different antigens (10% of the total number on the array) revealed a pattern of IgG antibody reactivity in the pre-CAD sera that discriminated between the mice resistant or susceptible to CAD with 100% sensitivity and 82% specificity (p-value$= 0.017$). Surprisingly, the set of IgG antibodies that was informative before CAD induction did not separate the resistant and susceptible groups after the onset of CAD; new antigens became critical for post-CAD repertoire discrimination. Thus, at least for a model disease, we found that present antibody repertoires can predict future disease, predictive and diagnostic repertoires can differ, and decisive information about immune system behavior can be mined by high-throughput technology.

Respiratory metabolism (paper IV)

NAD and NADP are two of the most used coenzymes in cellular metabolism. More than 500 known enzymes use NAD(P) to catalyze reduction-oxidation reactions reversibly[1]. Some of these are amongst the most abundant and well-studied enzymes participating in energy metabolism (glycolysis, the Krebs cycle, the Calvin cycle), biosynthesis, degradation, defense against oxidative damage, etc. NAD(P) is found in mitochondria, chloroplasts, peroxisomes, the cytosol and other cellular compartments.

Several methods are available for measuring total amounts of the two coenzymes (Stocchi et al., 1985, Agius et al., 2001), however no quantitative and simple method for distinguishing between free and bound versions has yet been developed. Consequently, in spite of the importance of NAD(P) in mitochondrial metabolism, little is known about the amount of free NAD(P)H, the parameter important for the interaction of the coenzyme with enzymes. Wakita et al. (1995) were unable to detect free NADH in rat liver mitochondria using fluorescence decay analysis. This implies that a small proportion of the total NADH is free. In contrast, the recent publication of Blinova et al. (2005) reported a high

[1] `www.chem.qmul.ac.uk/iubmb/enzyme` (Enzyme Nomenclature)

relative concentration of free NADH in pig heart mitochondria. This disagreement between the reports from different groups is a basis for confusion about the role of free NADH in mitochondrial metabolism.

The reduced and oxidized forms of the two coenzymes have distinct spectroscopic characteristics — NAD(P)H is fluorescent, while its oxidized form is not. Further, the fluorescence properties of bound and free NAD(P)H also differ. Using these fluorescent properties, and building on the qualitative work by Paul and Schneckenburger (1996), we developed a quantitative method for measuring both.

We use the developed method to (1) verify the presence of free NADH in mitochondria and (2) to investigate its role in the regulation of respiratory processes. It seems the free NADH concentration is kept constant under different metabolic conditions. To investigate possible explanations, we build a model that predicts a 2.5 to 3-fold weaker binding of NADH to mitochondrial protein in state 3 (substrate + ADP present) as compared to state 4 (substrate + ATP).

Chapter 1

Single cell technology

Animals and higher plants consist of a multitude of distinct tissue types. Furthermore, many of these tissues contains several different cell types, each performing specialized functions. In order to study these functions, single cell technology is needed.

First and foremost is the observation of single cells. The characteristic size of single cells is in the μm range, and consequently, the light microscope is a much-used tool. Of course, for the study of more detailed surface phenomena, higher resolution may be needed (as for example provided by scanning electron microscopy).

Having observed the cells; second is the identification of them. This may be relatively easy if the cells display obvious characteristics and are not moving, or may require more expertise and, for example, the assistance of computers with tracking software in the case of motile cells.

Third is the isolation of single cells; for culturing in growth media, or just for storage in suspension before further processing. For some tissues, cell (or protoplast) isolation can achieved by enzymatic digestion of the extracellular matrix (Birnbaum et al., 2003). However, it has been observed that the cells reacts to this, leading to changes in metabolite, protein, and mRNA levels (Celis et al., 1999). Another approach is to freeze or fix the tissue and then dissect the cells using manuel or more sophisticated laser-based techniques, for example laser capture microdissection, LCM (Emmert-Buck et al., 1996), and laser microdissection coupled to laser pressure catapulting, LMPC (Schütze and Lahr, 1998). In LCM, a plastic film is placed over a tissue section containing the cells of interest. At the spots where the cells are located, an infrared laser beam then melts the film, effectively making the cells stick to the film. Subsequent removal of the film also plucks the cells away from the tissue. In LMPC, an UV-A laser cuts out the cells of interest from the tissue. The excised cells are then ejected into a tube by a gas pressure cloud that is produced by a defocused laser pulse. In contrast to LCM, LMPC does not heat up the cells, and therefore avoids this source of potential damage to cellular components.

Fourth is the extraction of single cell parts or contents. If cells have been isolated prior to this step, disruption of the cell membranes (lysis) may be sufficient. If the cell is still located in the tissue, cell contents can be extracted by using a

glass microcapillary mounted on a micromanipulator to penetrate and aspirate the cell (Gjetting et al., 2004). Cell turgor is usually considerable and after penetration the cell contents are driven by it into the capillary tip. The capillary is then removed from the cell and the sample ejected into a buffer or some other storage medium. This technique for extraction is minimally invasive since only the cell in question is disrupted by the procedure (except in the case where the cell lies below other cells).

Fifth is the processing of cell parts or contents for measurement. Enough material must be collected or produced to reach a measurable threshold, for example by amplifying the mRNA content in the case of gene expression studies.

Three research examples of the use and development of such single cell technology are presented below. The first example (Sec. 1.1) concerns the processing of barley epidermal cells for subsequent hybridization on cDNA arrays. Here, a central issue is the amplification of the extracted mRNA. A theoretical treatment of this is given in Sec. 1.2, followed by an example of how the reliability of the amplification can be tested (Sec. 1.3). Sec. 1.4 describes how moving cells can be monitored and the resulting data be prepared for further analysis. Finally, Sec. 1.5 gives a few examples of technologies that measure mRNA, protein, and metabolite amounts in individuel cells, along with a discussion of the implications.

1.1 Research example: Identification of barley epidermal cells. Extraction and amplification of mRNA content (paper I)

Individual resistant and infected epidermal cells of susceptible barley were distinguished by microscopy of living barley leaves 18 h after powdery mildew inoculation (Fig. 1.1). The presence of a papilla (*P*) subtending the primary appressorial lobe (*L1*) and formation of a second lobe (*L2*) was used to recognise and select resistant cells (R) since a second lobe forms only when an effective host cell papilla prevents attempted penetration from the first. Infected cells (I) were recognised by the presence of a rudimentary fungal haustorium (*H*). The contents of individual resistant cells, infected cells, and uninoculated control cells were collected in microcapillaries. The mRNA in the samples was immediately captured on magnetic beads (coated with strings of T nucleotides to catch the corresponding polyA tails of the mRNA), purified, reverse transcribed to cDNA, and amplified using PCR (Polymerase Chain Reaction). During the amplification, radioactive nucleotides were incorporated in the amplified material. Using 35 amplification rounds ensured sufficient material for subsequent detection after hybridization to the 10k-array (see Sec. 2.1.4).

1.2 Amplification of DNA

To increase the amount of cDNA to detectable levels, in general, two different amplification strategies can be used (Winter et al., 2002). Linear amplification

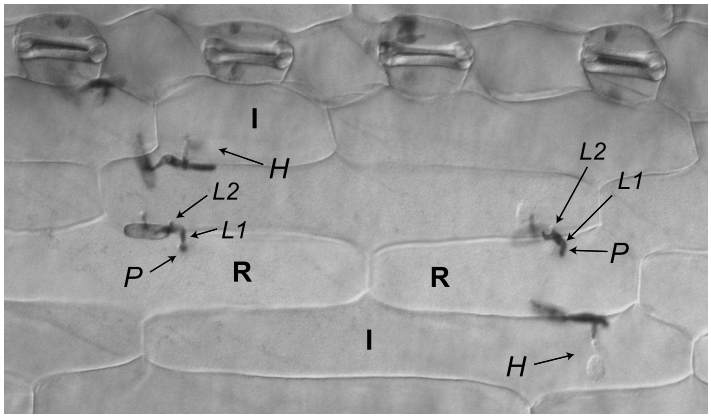

Figure 1.1: Microscope image of the epidermal layer of a barley leaf inoculated with powdery mildew. I: Infected cell, R: Resistant cell, *H*: Haustoria, *P*: Papilla, *L1*: Primary lobe, *L2*: Secondary lobe.

in the form of in-vitro transcription, or exponential amplification, as is achieved when using PCR.

Linear amplification

For in-vitro transcription, the enzyme RNA polymerase produces new cRNA transcripts from a population of cDNA templates. Each given cDNA template is transcribed many times over, producing several identical cRNA sequences. Let $N_i(0)$ be the number of cDNA templates of a given type, i, present in the sample to begin with. Let λ_i be the rate of amplification of this template. After a time t, the expected number of cRNA sequences transcribed, denoted $N_i(t)$ will be

$$N_i(t) = N_i(0)\lambda_i t$$

Events (successful amplification) that occur at a definite average rate (λ_i) is governed by the Poisson distribution. Consequently, the variance is given by

$$\sigma_i^2(t) = N_i(t) = N_i(0)\lambda_i t$$

and the coefficient of variation (the relative variation) by

$$\mathrm{cv}_i^{\mathrm{linear}}(t) = \frac{\sigma_i(t)}{N_i(t)} = \frac{1}{\sqrt{N_i(0)\lambda_i t}} \tag{1.1}$$

Exponential amplification

PCR is based on a thermostable form of the DNA polymerase enzyme. For a given sequence of cDNA, one needs first to design primers that flank this sequence. The reaction works in cycles, where each cycle comprises: (1) Denaturing of the DNA into two single strands by heating, (2) cooling the solution

which catalyzes primer annealing to each of the single strands, (3) medium heating which activates the DNA polymerase to begin producing a complementary strand to each single strand, starting from the primer, (4) denaturing this newly formed double-stranded DNA by heating, the same as in (1), etc. Around 30 cycles is common. Compared to the linear amplification method, this method makes copies of copies. Therefore, it very quickly generates large amounts of amplified material. However, the variation in the final amount of amplified material differs markedly from the linear amplification case, as the following calculation show. Let $M_i(0)$ be the number of cDNA templates of a given type, i, present in the sample to begin with. Let γ_i be the probability of successful amplification of this template in a given round. After k rounds, the expected number of sequences, denoted $M_i(k)$ will be

$$M_i(k) = M_i(0)(1+\gamma_i)^k$$

That this is so can be quickly seen when remembering that the series $1, 2, 4, 8, \ldots$, which doubles for each successive element, can be written as 2^k. Compare with the above equation when setting γ_i equal to 1. The equation shows that the amount of amplified material scales exponentially with the number of rounds, hence the name exponential amplification. Further calculation show that the variance can be written

$$\sigma_i^2(k) = M_i(0)(1-\gamma_i)\Big((1+\gamma_i)^{2k-1} - (1+\gamma_i)^{k-1}\Big)$$

and thus the coefficient of variation as

$$\mathrm{cv}_i^{\exp}(k) = \frac{\sqrt{(1-\gamma_i)\Big((1+\gamma_i)^{2k-1} - (1+\gamma_i)^{k-1}\Big)}}{\sqrt{M_i(0)}(1+\gamma_i)^k} \tag{1.2}$$

Several papers have been published on the mathematics of PCR. Two papers that cover the equations mentioned here are Stolovitzky and Cecchi (1996) and Shaw (2002).

Comparison of linear and exponential amplification

In Fig. 1.2 is shown the coefficients of variation for linear (A) and exponential (B) amplification. Note that whereas for linear amplification the relative variation decreases to zero in the limit of $t \to \infty$ (Fig. 1.2A) and (1.1), it reaches an asymptotic value not equal to zero for exponential amplification (Fig. 1.2B) and (1.2). This is very unfortunate since only the PCR amplification method can amplify the small amounts of cDNA converted from extracted mRNA in reasonable time. However, as the equations show, when amplifying a pool of different cDNA templates, the final amounts of each template will vary. Thus, the relative levels of each type of mRNA are not necessarily preserved through the amplification. Though, as is seen in Fig. 1.2B, the larger the amount of initial copies of each template, the smaller the cv. That is, it is absolutely necessary to pool cells until a sufficiently small cv is reached.

In these derivations it was assumed that λ_i did not change as a function of the number of rounds. In reality, the experimental protocol needs to be carefully

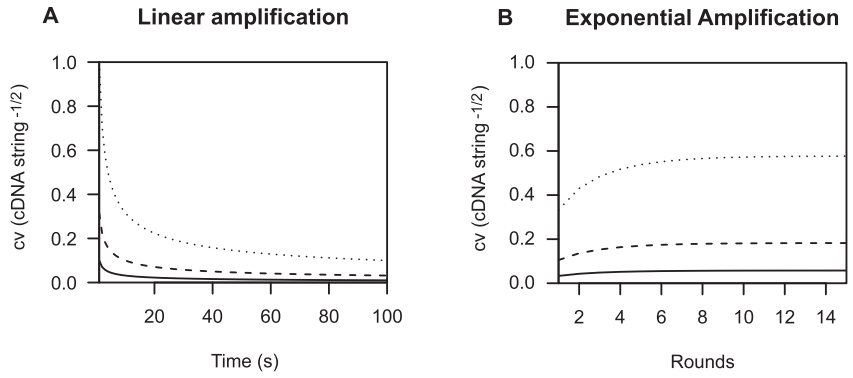

Figure 1.2: Coefficient of variation (cv) as a function of (A) time in linear amplification and (B) rounds in exponential amplification. Solid lines: 100 cDNA strings, dashed lines: 10 cDNA strings, dotted lines: 1 cDNA string, as starting material.

prepared in order to insure this, and insure that it as close to one as possible. Several groups have developed exponential amplification methods that works at the lower limit of the amount of starting material needed to conserve the relative abundances of different transcripts (Karrer et al., 1995, Gallagher et al., 2001, Brandt et al., 2002, Iscove et al., 2002). For the barley epidermal cells described in Sec. 1.1 we find that mRNA pooled from around 20 cells provides a sufficient amount for expression analysis. This claim is substantiated by two observations: (1) By an independent test described below, and (2) by comparing the cDNA array data obtained using this material to other barley/powdery mildew studies where the mRNA was not exponentially amplified (described in Sec. 3.3.6).

1.3 Research example: Test of exponential amplification of single-cell material (paper I)

The test of reproducibility of the exponential amplification is illustrated in Fig. 1.3 and described below. We extracted and pooled material from 30 infected and 30 control cells. One third of the material from each of these pools was combined to create a new, mixed sample. This sample thus represented a biological average of the samples containing pure infected and control material (Fig. 1.3A). The resulting three samples, each equivalent to 20 cells' contents, were PCR amplified and hybridized to small dotblot arrays spotted with 91 ESTs isolated from powdery mildew-infected barley cells (concerning ESTs, see Sec. 2.1.1). We know that these 91 genes covers a reasonable part of the total range of expression values in the cell (as measured on the 10k-array, see Sec. 2.1.4), see Fig. 1.4A, and that about one third are clearly differentially expressed between the two cases (26 genes are more than two-fold up- or down-regulated). If the amplification step preserves the original abundance relationships, the expression of each gene measured in the mixed sample should be

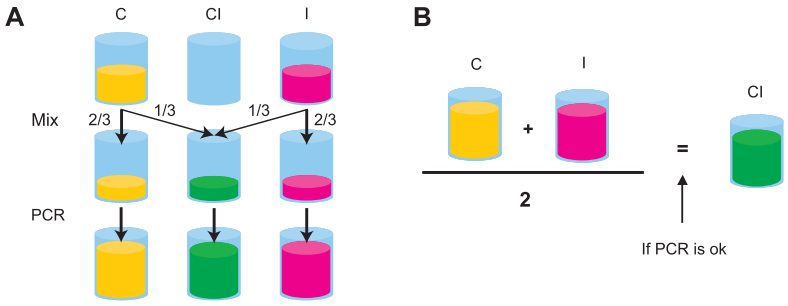

Figure 1.3: (A) One third of a control sample, C, and one third of an infected sample, I, is mixed together. The resulting three samples are PCR amplified. (B) If the PCR amplification steps preserves the relative expression of each gene, the expressions measured in the mixed sample should equal the average of what is measured in the control and infected samples.

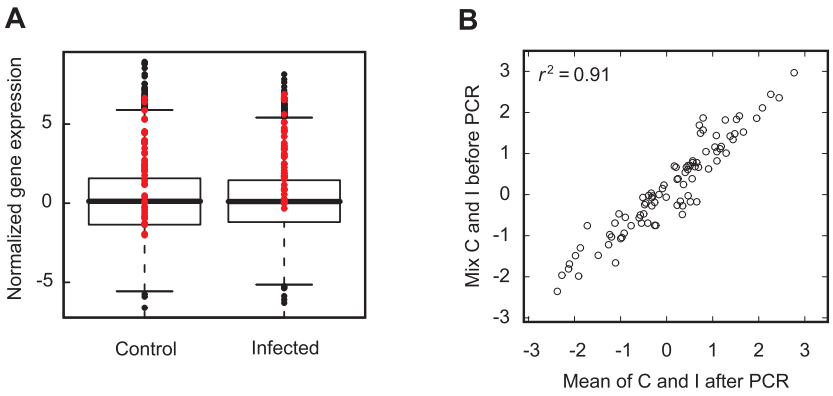

Figure 1.4: (A) Coverage of expression range. Boxplots: total range of expression values on the 10k-array. Red dots: 10k-array expression values of the 91 genes spotted on dotblot arrays. (B) Validation of PCR amplification. The average of the expression values in the infected and control samples as measured on the dotblot arrays (x-axis) are plotted against the expression values in the mix of infected and control sample (y-axis).

similar to the average expression in infected and control material (Fig. 1.3B). As is seen in Fig. 1.4B, this is clearly the case for the entire range of expressions intensities. This demonstrated that different genes having different expression patterns and levels were amplified to a reasonably equal degree (coefficient of determination $r^2 = 0.91$) and that the protocol developed for PCR maintained relative transcript abundance levels.

1.4 Research example: Motile human keratinocyte- and dermal fibroblast cells (paper II)

In the following, techniques for observation and basic data handling of motile cells are described.

The motility data were obtained from HaCaT cells on substrata of molecular- and fibrillar collagen (Gadegaard et al., 2003), and from NHDF cells on the same two collagen substrata, as well as on glass and tissue culture grade polystyrene, TCPS, all at 37°C. HaCaT cells did not show measurable motility on glass and TCPS during the first 24 h of the experiment.

A microscope stage was converted to a temperature-controlled incubation chamber with a computer-controlled step motor for parallel monitoring of several sites and samples under equal conditions. Cells were monitored in an inverted phase contrast microscope with autofocusing controlled by dedicated software. A digital camera recorded time-lapse movies with intervals Δt of 15 min for 24 h.

Only "healthy" cells with full spreading on the substratum and long presence in the view field were selected from recorded image sequences. One pixel in each cell nucleus was marked manually as the cell's coordinate. Cell populations were dilute, so direct interactions between motile cells were rare. When they occurred, both cells' trajectories were excluded from the statistics. Cell trajectories crossing each other were less rare. No effect of this could be discerned. Data for cell coordinates were corrected for drift, and single cell trajectories were calculated from position and time data. A number of cells had to be removed from the statistics because they did not move. Visual inspection of trajectories would reveal immobile cells. From a number of such inspections, it was found that they could be excluded automatically, by excluding cells with a root mean square deviation (RMSD), or standard deviation, that never exceeded 15 mm in the case of HaCaT cells and 20 mm in the case of NHDF cells.

For each cell trajectory $\boldsymbol{r}(t)$; the positions $\boldsymbol{r}_j = \boldsymbol{r}(t_j)$, $t_j = j\Delta t$, $j = 1, 2, 3, \ldots$ were recorded, and the velocities $\boldsymbol{v}_j = (\boldsymbol{r}_j - \boldsymbol{r}_{j-1})/\Delta t$ were calculated. From these velocities, the accelerations $\boldsymbol{a}_j = (\boldsymbol{v}_j - \boldsymbol{v}_{j-1})/\Delta t$ were formed. Since surfaces were manufactured to be homogenous and isotropic to cells, and the cells' environment was kept constant in time, we tested and found velocity distributions consistent with spatial homogeneity and isotropy, as well as with temporal invariance. Consequently, when computing any ensemble average, we also averaged this quantity over time, space, and direction to improve statistics.

1.5 Examples of expression measurements in individual cells

Even though the method described in Sec. 1.1 samples individual cells of the same type, as confirmed by microscopic inspection, it still pools the contents of several cells in order to reach measurable levels of mRNA. For the biological questions asked this is not problematic. The idea is to understand the general (or average) response to infection or resistance, and we do not ask why some epidermal cells get infected and others do not. However, say that one could, in fact, measure the transcript profile in individual epidermal cells. Could it be, that even though epidermal cells are similar in the microscope (before inoculation), their transcript, protein, and metabolite levels are not identical — maybe even very different in some respects? If so, do these differences determine the outcome of the attempt to infect the cell? This is an open question. Relating to this discussion, in Sec. 3.3.8, we show that genetically identical mice display clearly different antibody profiles. And furthermore, that these profiles can be used to predict the response to treatment. Now, barley epidermal cells and mice are clearly different, but the biological mechanism might be the same. To conclude this chapter and expand on the above stated question, three recently developed techniques for measuring the amounts of mRNA, protein, or selected metabolites in individual cells, are presented.

Gene Expression

Using fluorescence in situ hybridization (FISH) and a multiplex probe design, Levsky et al. (2002) analyzed the gene expression of 11 genes at 14 time points over a 90 min period in individual normal fibroblast cells. FISH is a well established method for detecting DNA or mRNA in situ by using labelled oligonucleotide probes (Levsky and Singer, 2003). The introduction of amino-allyl modified bases allowed the chemical synthesis of multiple-labeled fluorescent hybridization probes (Femino et al., 1998). And this, in turn paved the way for applying multicolor FISH to visualization of multiple RNA species simultaneously (Levsky et al., 2002). More specifically, it is the synthesis of mRNA that is monitored by visualizing specific sites of transcription. The results show that there is considerable cell-to-cell variability, even for synchronized cell-cultures. It seems that each gene can be thought of as a stochastic variable, subject to various types of noise, as modelled by Elowitz et al. (2002).

Protein amounts

Hu et al. (2004) have developed a two-dimensional capillary electrophoresis method for studying the amounts of protein in single mammalian cells. These cells, containing as little as 50 pg of total protein, are aspirated into a 50 μm-diameter capillary by a brief pressure difference. The cell is then lysed inside the capillary to release its complement of protein. The proteins are labelled with a fluorescent tag and then separated inside the same capillary by their molecular weight (using capillary sieving electrophoresis, which is the capillary version of conventional polyacrylamide gel electrophoresis). To separate the proteins

further, successive fractions are transferred to a second capillary. Applying an electric field across this capillary separates proteins based on their interaction with the surfactant in a micellar electrophoretic separation. Repeating this process about 100 times under computer control comprehensively separates the protein content of the single cell. The minute amounts of protein that result are detected by laser-induced fluorescence. Using this method, cell-to-cell variation of protein profiles in cultured breast cancer cells was observed — both before and after treatment.

Metabolite amounts

Using the same setup with microcapillaries described in the beginning of this chapter and in Sec. 1.1, one can from the extracted material measure metabolites instead of mRNA levels. From individual cell samples containing as little as $10 - 100\,\mathrm{pL}$, the metabolites and compounds nitrate, glucose, fructose, sucrose, fructans, malate, mannitol, and glucosinolates have been successfully measured (Tomos and Sharrock, 2001). The measurement technique is based on fluorescent microscope photometry linked to an enzymatic redox reaction involving NAD(P)H. For example, using this method it was shown that the malate concentration is approximately six times lower in barly epidermal cells lying close to stomata (breathing pores on the leaf surface) compared to cells lying further away from them (Fricke et al., 1995). This observation makes sense since malate plays a signalling role in the stomatal guard cells, and the transport of malate from nearby epidermal cells thus needs to be tightly controlled.

Chapter 2

Omics methods and technologies

Consider the large number of different tasks that a cell must perform: Synthesis, degradation, and exchange of biomolecules in a number of metabolic processes, maintenance and repair of its constituents, and in general helping the organism to adapt to the environment (for example by learning, in the case of nerve cells), to mention a few. Cells can respond to external chemical signals by for example growing and dividing, differentiating into specialized cell types, initiating or halting secretion, starting to proliferate, leaving their place in the extracellular matrix, or initiating apoptosis. Other external stimuli can also elicit specific responses: attacks on the plant cell wall by fungi mobilize a myriad of defense responses such as cell wall strengthening, and the motility patterns of moving cells are distinctly different when moving on different surfaces — just to name two examples central to the research for this thesis. Also, cellular dysfunction may result in serious diseases, for example cancer (uncontrolled cell division) or Huntington's disease (continuous aggregation of huntingtin molecules in neuronal cells).

To aid in the description (and perhaps understanding) of how the cell performs all these functions, one can divide the molecules in the cell into large groups, depending on their type and function. These groups are: The genome, transcriptome, proteome, and metabolome. The genome is the complete DNA sequence found in the cell. This includes both genes and non-coding sequences. The suffix "ome" is the Greek word for "all", "every" or "complete". When using the suffix "omics" instead, it is the corresponding field of study that is referred to. Consequently, genomics is the study of an organism's genome and the use of the genes. The mRNA transcribed from the genome is termed the transcriptome. In contrast to the genome, which is a rather constant entity, the transcriptome differs from cell to cell and is constantly changing in response to both external and internal stimuli. Translating the transcriptome we have the proteome (the entirety of proteins), the large-scale study of their structure and function of which is the field of proteomics. Finally, there is the metabolome, which is the complete set of small-molecule metabolites (such as metabolic intermediates, hormones and other signalling molecules, and secondary metabolites). The me-

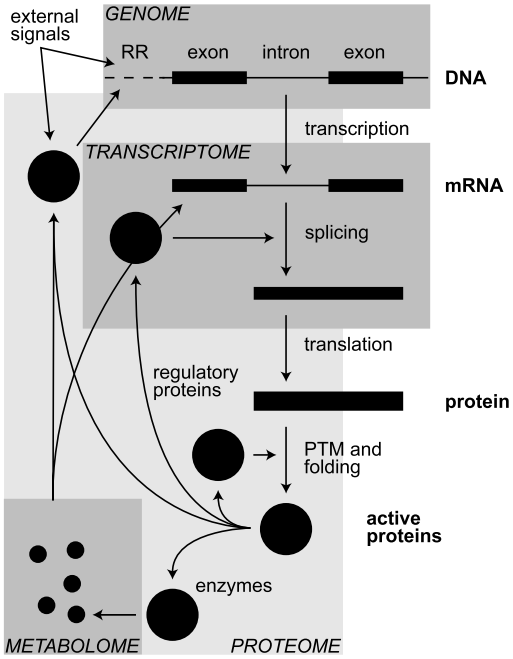

Figure 2.1: Transcriptional and translational mechanisms and control points for regulation. RR: regulatory region, PTM: post-translational modification.

tabolome might be the most dynamic of all the omes, changing from second to second in response to the cell's needs.

The robust and adaptable behavior of cells ultimately depends on the operation of complex regulatory biochemical networks spanning the different omes. Even networks performing "simple" functions are often remarkably complex, involving proteins, genes, and metabolites in multiple interlocked regulatory loops. To appreciate this, below, a brief overview of gene expression and its regulation is presented (Stetter et al., 2003). Simply put, proteins are produced from genes; they interact with each other and with metabolites but also act back onto the DNA where they regulate the transcription of mRNA and thus the production of other proteins, see Fig. 2.1. Starting from the top to the right in this figure, a gene in the genome is transcribed into mRNA. The non-coding intron regions are spliced out, and the exons combined[1]. The spliced mRNA is translated into an amino-acid chain, the protein. Proteins may be chemically modified (post-translational modifications) by enzymes before folding into their final shape. They interact with each other, with metabolites, or with parts of the genome (transcription factors) or transcriptome (proteins involved in mRNA splicing). Metabolites may interact with regulatory proteins or mRNA (Kaempfer, 2003). Finally, external signals can interact with the system as transcription factors or by modifying transcription factors.

[1]Exons can be combined in different ways. This is one way to produce several different proteins from the same gene sequence.

In the next three sections, experimental methods and technology central for studying each of the omes and their possible interactions are presented.

2.1 Genomics and transcriptomics

Central to genomics and transcriptomics is knowing the order of nucleotides in DNA and RNA. Determining this order is called *sequencing*, and will be discussed in Sec. 2.1.1. In Sec. 2.1.2, methods for large-scale measurements and analysis of gene expression are presented. A key technology is *microarrays*, which are described in some detail. The preprocessing of microarray data is specifically discussed in Sec. 2.1.3. Finally, Sec. 2.1.4 presents an example of the development of a spotted cDNA microarray.

2.1.1 Sequencing

In 1977, two robust techniques for rapid DNA sequencing were introduced: The enzymatic chain termination method by Sanger et al. (1977), and the chemical degradation method by Maxam and Gilbert (1977). Since then sequencing techniques have been improved and automated significantly by e.g. specifically engineered enzymes and dyes and the use of robots (Sterky and Lundeberg, 2000). To date (March 2006), 355 genomes have been sequenced, divided into 41 eukaryal, 26 archaeal, and 288 bacterial genomes (Liolios et al., 2006)[2]. Furthermore, in the EST database dbEST[3] (Boguski et al., 1993), more than 32 million EST sequences (see below) have been submitted from 1020 different genomes. Thus, today there is a vast amount of sequence information stored in public repositories. The knowledge contained in all these sequences is a cornerstone of high-throughput modern biology. The fundamental sequencing technique that has spawned this development will therefore be described in some detail below.

The basic sequencing reaction, based on the enzymatic chain termination method, consists of (1) the DNA sequence one wants to sequence (the DNA template), (2) a primer sequence, from which the reaction can be initiated, (3) each of the four nucleotides (dATP, dCTP, dGTP, and dTTP), (4) the DNA polymerase enzyme, which, starting from where the primer is attached to the template string, repetitively inserts complementary nucleotides and thus builds up a new DNA string, complementary to the template, and (5) limiting amounts of each of the four dideoxy versions of the nucleotides (ddATP, ddCTP, ddGTP, and ddTTP). Whenever the DNA polymerase tries to insert one of these into the growing chain, the reaction will stop and the chain terminate. Since there are only few of the dideoxy nucleotides in solution compared to the normal ones, there is only a small chance that the growing chain will terminate at any given position. In one reaction chamber there are multiple identical DNA templates and multiple primers, and thus, a nested set of DNA fragments, each terminating at a specific base, are generated. Since each of the fragments are tagged with a fluorophore that reveals which kind of dideoxy nucleotide that terminated the reaction, when separating the fragments according to size by running them

[2] `www.genomesonline.org`
[3] `www.ncbi.nlm.nih.gov/dbEST/`

through a gel with an electric field applied, one can deduce the sequence of nucleotides in the DNA template by measuring the fluorescent tag each fragment has as it exits the gel (the smallest fragment containing only a single nucleotide exits the gel first, the next fragment contain two nucletides, then three, and so on).

With current technology, DNA strings of maximally around 800 nucleotides can be sequenced in one reaction. Therefore, when sequencing genomic DNA it has to be broken up into smaller pieces first. A more direct access to genes, which can be difficult to locate in genomic DNA, is to reverse transcribe mRNA to cDNA, and then sequence the cDNA. In this approach, one usually first creates a cDNA library by inserting the cDNA into viral or plasmid vectors, denoted cDNA clones. This enables rapid replication of the cDNA, which can then be sequenced. These sequences are called expressed sequence tags (ESTs).

2.1.2 Methods for large-scale gene expression analysis

Traditionally, Northern or Southern blots have been used to measure gene expression. Here, the gene of interest in the form of a labelled RNA or DNA string is incubated with RNA or DNA from the cells of interest. The cellular RNA or DNA has been separated according to size in a gel, and if a complementary sequence is present, the labelled string will hybridize to it. After washing, this can be detected as a band on a x-ray film. The method is not easily extended to hundreds of genes, and is only semiquantitative.

Two other approaches, both based on PCR, are differential display (Liang and Pardee, 1992) and real time quantitative PCR (Heid et al., 1996). The latter method has been used in a 384-well-plate based system, thus measuring the expression of 384 genes simultaneously[4]. Although these methods measure the expression of several, even hundreds of genes, they are still not suited for truly large-scale gene expression measurements.

Yet another approach to expression profiling is based on sequencing. Here, random sequencing of cDNA generated from the cells of interest are followed by counting the number of times each type of cDNA string is sequenced. These counts reflect the expression of the corresponding gene. The method is denoted serial analysis of gene expression, SAGE (Velculescu et al., 1995).

DNA microarrays is by far the method that is most commonly used today for large-scale gene expression analysis. Microarray technology is not limited to genomics and transcriptomics. In general, a microarray consists of a solid surface onto which a large number (from around 1000 to 50000) of molecules with specific binding properties have been chemically attached. The idea is that these molecules can act as probes for molecules in a solution (effectively a reverse Northern blot in the case of gene expression). Depending on the type of probe, several types of microarrays have been devised. The first microarray created was a DNA microarray (Schena et al., 1995), in which labelled DNA molecules in a solution hybridizes to complementary DNA on the array. Other examples of microarrays are protein microarrays, tissue microarray, or antigen

[4] products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=600759&tab=DetailInfo

microarrays, in which antigen probes on the array will bind specific antibodies in a serum sample (see Sec. 2.2.3). In the following, DNA microarrays will be described in some detail (and will simply be referred to as microarrays).

There are two main technologies for making microarrays: robotic spotting and in situ synthesis. For the spotted microarrays, the solid surface is usually either a glass slide or a nylon membrane. A spotting robot is used to put the probes on the surface, the probe material being for example PCR products or presynthesized oligonucleotides. To aid in the subsequent detection, the probes are arranged in an orderly fashion on the array, for example equidistantly in rows and columns.

For the in situ synthesized microarrays, oligonucleotides are built up base by base on the surface of a glass slide. Each base added to the oligonucleotide on the array has a protective group to prevent the addition of more than one base during each round of synthesis. The protective group is then removed with either light or acid (depending on the type of protective group used) before the next round of synthesis. So, in each addition cycle, the following steps are performed: (1) The photo- or chemolabile group on the oligonulceotides in a specific region are deprotected using either light or acid. (2) Nucleotides (either A, C, T or G) are flushed over the array, binding to the exposed oligonucleotides. (3) The remaining free nucleotides are removed. And (4), the uncoupled nucleotides are capped by chemically inactivating the reactive hydroxyl group.

When using light as the method for deprotection, two approaches are currently used to direct the light to the relevant regions of the array: (1) By using a mask that allows light to pass to some regions but not to others (each round of synthesis then requires a different mask). This is the method used by Affymetrix[5]. And (2) by having an array of computer-controlled micromirrors, which can be adjusted to direct light on the appropriate regions of the array. This is the method used by Nimblegen[6] and Febit[7].

Instead of using light for deprotection, one can also use chemical deprotection with synthesis via inkjet technology. At each step of the synthesis, droplets of the appropriate base are fired onto the desired spot on the glass slide using what is essentially an inkjet printer. This is the method used by Rosetta[8], Agilent[9] and Oxford Gene Technology[10]. The advantage of the micromirror and inkjet technology over the use of masks for deprotection is that the array can contain exactly the oligonucleotides the array-maker wishes. That is, it is highly flexible. However when using masks, once the masks have been made, a large number of identical arrays can be produced efficiently. The in situ synthesized microarrays have a lower variation between technical replicates than the spotted microarrays, but are generally more expensive. A number of less conventional microarray technologies are also being developed. For example, the use of semiconductor chips as solid support. The oligonucleotides are then build up by activating specific electrodes that deprotects the oligonucleotides by an electrochemical

[5] www.affymetrix.com
[6] www.nimblegen.com
[7] www.febit.de
[8] www.rii.com
[9] www.chem.agilent.com
[10] www.ogt.co.uk

reaction (CombiMatrix[11] and Nanogen[12]).

Microarrays can also be used to detect mutations or polymorphisms in specific genes. On these arrays, groups of individual sequences differ only from each other by a single or a few nucleotides. Since these microarrays genotype whole genomes by measuring Single Nucleotide Polymorphisms (SNPs) across the genome, they are called SNP- or genotyping microarrays (Kidgell and Winzeler, 2005). By combining SNP-microarray analysis with knowledge about haplotypes (sets of nearby SNPs on the same chromosome that are inherited in blocks), as for example recently published for the human genome (The International HapMap Consortium, 2005), genes involved in inherited diseases can be comprehensively identified.

2.1.3 Using microarrays

Several steps are needed before data from microarray experiments can be analyzed. The slides need to be scanned and the resulting images segmented into spot and background regions. The background intensity must be removed and an estimate of the intensity of the material hybridized to each spot calculated. Finally, all spot intensities on each of the microarrays in the experiment must be normalized to adjust for systematic effects. These normalized intensities then hopefully provide reasonable estimates of the expression of the corresponding genes. The individual preprocessing steps are discussed in more detail below. In general, for each step, a range of methods are available. Diagnostics that can be performed to help select the optimal method are described and also, in the end of this section, a general scheme for systematically testing which combination of preprocessing steps is the optimal for a given technology and experiment is presented.

Scanning

The hybridized material is labelled with either a fluorescent or radioactive tag. When using radioactive tags, such as the beta-emitting radioisotope ^{33}P incorporated into the nucleotide triphosphate molecules, the detection process, denoted phosporimaging, consists of two steps. (1) The radiation from each spot is detected using a phosphor storage screen. The detection layer in such screens consists of BaFBr:Eu crystals. When an europium (Eu) electron is excited by the radiation, it is trapped in the crystal lattice structure producing a stable BaFBr:Eu^{3+} crystal. (2) The screen is then scanned with a laser. When the laser hits the crystals containing excited electrons, they are released and emits fluorescent light upon returning to the ground state. This emitted light is detected by a photomultiplier tube and recorded (e.g. in a TIFF file).

When using fluorescent tags, the array is scanned directly and the emitted light from each spot detected and recorded. Fluorescence has the advantage that samples labelled with different fluorescent tags can be mixed and hybridized to the same array (two-color cDNA arrays). Also, effects such as overshining (see

[11] www.combimatrix.com
[12] www.nanogen.com

below) are avoided. On in situ synthesized oligonucleotide microarrays only one type of fluorescent tag is used.

The resolution of the resulting images differs depending on the type of array and the technology used. For glass slides spotted with fluorescent material, spot diameters usually range from 25–50 μm and the resolution is around 3 μm per pixel. Radioactive spots on nylon membranes are usually larger (around 0.5 mm), and are also scanned at lower resolution (around 50 μm). In both cases, each spot is contained in around 100 pixels.

During the scanning of microarrays hybridized with fluorescent material, noise can influence the measurements, sometimes quite drastically, producing outlier pixels with much higher or lower intensities than the pixels around them. Usually, the microarray is only scanned once. However, if the array is scanned several times, this could potentially be used to detect such outlier pixels and remove them (Romualdi et al., 2002). Otherwise, noisy pixels, etc. are usually handled in the segmentation phase.

Segmentation

In the following, only segmentation of images from spotted microarrays are described. For in situ synthesized oligonucleotide microarrays, the individual oligonucleotides are located with such precision, and so close together, that that there is no need for the methods described below.

Having scanned the spotted microarray and obtained an image of it, next is the segmentation of the image into spot and background regions. Ideally, only the spots should emit light and thus be easily separated from the black background. In reality however, the background also binds some of the labelled material. In addition, dust particles, contaminations, smears of spot material over large areas, etc. are also often seen. Therefore, algorithms are needed that can classify pixels into spot and background. The first step in segmentation is to place a grid over the image. The grid is adjusted so that each grid square contain one spot. Methods to automate this process have been proposed (Jung and Cho, 2002, Galinsky, 2003), however they have yet to come in widespread use.

With the grid in place, several different algorithms have been developed that attempt to separate spot from background. A simple approach is the fixed circle algorithm. Here, a circle with a predetermined diameter is placed in the grid square in a manner that maximizes the difference between the pixel intensities inside the circle and outside the circle. One can extend this method and use two circles, the one larger than the other. The pixels between the two circles, corresponding to a transition area, are discarded, and this usually gives better estimates for spot and background intensities. Often spot sizes are not uniform. To accommodate this one can let the algorithm optimize both the position and the diameter of the circles (adaptive shape algorithm). Sometimes spot shapes are not even circular, or artifacts such as smears from other spots partly overlay the spot. To handle these situations, the histogram method or the seeded region growing algorithm have been proposed. The histogram method removes the lowest and highest intensity pixels found in the spot and background regions respectively, according to predetermined thresholds. The seeded region growing algorithm takes another approach. First, a set of pixels called seeds have to be

specified (for example, the pixels in the center of the grid square are the spot seeds and those on the grid intersections are the background seed). Then, the seeds are grown by defining a rule concerning the adoption of pixels adjacent to seeds. Usually the adjacent pixel with the intensity that is closest to the average of the seed pixel intensities is adopted as a seed. Finally, a rule for stopping the process has to be defined.

The choice of segmentation method can significantly influence the quality of the data. Studies that compare methods generally show that histogram- and seeded region-type methods are superior to the fixed or adaptive circle methods (Yang et al., 2002, Ahmed et al., 2004) (although the first paper puts more emphasis on choosing an appropriate background correction scheme, as discussed below). Hybrid methods that implement both histogram and seeded region growing have recently been reported to give even better results (Rahnenführer and Bozinov, 2004).

In the end, the goal is to develop a segmentation pipeline that performs all the steps automatically, and attempts at this have been proposed (Jain et al., 2002, Angulo and Serra, 2003). In general, however, different microarrays pose different problems, and methods (or combinations of methods) have yet to be developed that clearly perform well for all of them. Until such a time, informed manual intervention in the process is definitely advantageous when one wants to ensure high quality data.

Summary values and quality measures

Having segmented the pixels in each grid square into spot and background pixels, summary values can calculated. For example, to represent the intensity in a spot, the mean or median of all spot pixels can be used. For high intensity spots, a common problem is that a fraction of the spot pixels are saturating. That is, for 16-bit TIFF images, their intensity is $2^{16} - 1$. In reality, these pixels should have had higher values, but due to the settings of the scanner, their intensity exceeds what can be measured and stored. By calculating the standard deviation of spot pixels in each spot, such saturation can be detected since the spot standard deviation in these cases will be low. Instead of discarding such spots, or use the biased summary values, one could replace the saturating pixels with an estimate of their real values. Such estimates, based on censored regression (Dodd et al., 2004), or spot shape modelling (Ekstrøm et al., 2004), have recently been proposed.

Other summary values, that function as quality measures, can be calculated, for example spot size, variation in background pixels, circularity, and signal-to-noise ratio (difference between spot and background intensity measured in units of the spot standard deviation). These quality measures can be used to evaluate the individual spots, perhaps discarding some, or to adjust their spot intensities (Wit and McClure, 2003). One could also calculate an overall measure of quality which can then be used to weigh intensities in subsequent analyses (Wang et al., 2001, Brown et al., 2001, Bylesjö et al., 2005).

Overshining

Specifically for radioactively labelled arrays, it is important to correct for overshining. Since there is a small distance between the nylon membrane and the phosphor screen (due to the plastic wrapping of the membrane and the overcoat on the screen), and because radiation emits equally in all directions, the width of the intensity profile detected on the screen will depend on the intensity of radiation in the spot on the array. Therefore, the more radiation a spot emits, the larger the spot will be. Also, much of the intensity measured in background regions, and often also some of the intensity in the spots themselves, contains significant contributions from radiation emitted from neighbouring spots. Three approaches to solve this problem have been suggested. Therneau et al. (2002) uses an image deconvolution method to sharpen the images before segmentation. Machl et al. (2002) suggest to take advantage of replicate spots to calculate the average percentage that each spot overshines into neighboring spots, and then use this to correct spot intensities. And finally, Lopez et al. (2004) has calculated the spot profile one would detect from radioactive material uniformly distributed in a disk, and uses this to deduce the "real" intensity in each spot. In reality, however, spot material is often irregularly distributed and tends to aggregate in donut shapes.

Background correction

If the intensity measured in a spot, I_{tot}, is simply the intensity of the background in that spot, I_{bgr}, plus the intensity from the spotted material, I_{mat}, one could estimate the background intensity in the spot from the intensity in the background region around the spot, $I_{\mathrm{bgr}} \equiv I_{\mathrm{aro}}$, and subtract this number from the total spot intensity measured. This difference would then represent the intensity of the spotted material, $I_{\mathrm{mat}} = I_{\mathrm{tot}} - I_{\mathrm{aro}}$. Depending on the quality of the estimates of I_{tot} and I_{aro}, sometimes this is not the case. For example, when subtracting an overestimated background value from a low intensity spot, one might get a value below zero. And also, the variance between replicate spots for which the background has been subtracted may increase rather drastically (the difference of two noisy numbers). Several approaches have been suggested to avoid this. These range from the very simple ones where spots with negative values are removed from the dataset, to log-linear interpolation methods (Edwards, 2003), and methods that fit specifically chosen distributions to the foreground intensities using the background intensities as a covariate (Smyth, 2005).

A simple way to evaluate the performance of the segmentation and background correction is to plot the estimates of the background intensities against the background corrected spot intensities. If the background corrected spot intensities correctly represent the intensities from the spotted material only, there should be no dependence on the background intensities visible in the plot. That is, the points should scatter randomly and should not display a trend such as scattering around a line with positive slope.

Normalization

In order to be able to compare samples hybridized to different arrays, the only differences between samples should be the ones that are biologically interesting. That is, the differences that characterize the response to different treatments, etc. However, unequal quantities of starting mRNA, differences in labelling or detection efficiencies, or other array- or sample-systematic biases are the norm, and these need to be removed before comparisons can be made. This process is called normalization of the data. There are several approaches for normalizing microarray data. The simplest being the division (or subtraction if logarithms have been taken, see Sec. 3.3.3) of a constant such that the mean-, or median expression, is equal to 1. This assumes that the differences between arrays affect all spots on a given array equally. Often this is not the case, which can be seen e.g. by plotting all intensities on one array against all intensities on another array. If the slope of this plot is different from 1, the artifacts perturbing the measurements are intensity dependent. In this case, normalization methods such as quantiles (Bolstad et al., 2003), lowess (Yang et al., 2002), or qspline (Workman et al., 2002) may be used. As an example, the qspline method is described in more detail below.

By x_{ki} is meant the intensity of spot k in array i. The geometric mean (v_k) of each spot is calculated as

$$v_k = \left(\prod_{i=1}^{I} x_{ki} \right)^{1/I}$$

This forms the distribution which all the arrays are scaled against. The geometric mean is used instead of the mean because the distribution of x's is asymmetric, having a long tail for high x. This is also the case for log-transformed intensities. From the distribution of the v_k's and the distribution of the x_{ki}'s in each array, the 100 percentiles are calculated and placed in the vectors $\mathbf{q}_v = (q_{1v}, q_{2v}, \ldots, q_{100v})$ (for v) and $\mathbf{q}_i = (q_{1i}, q_{2i}, \ldots, q_{100i})$ (for each of the arrays). A cubic spline function, $s_i(x)$, is interpolated between the points $\{(q_{1i}, q_{1v}), (q_{2i}, q_{2v}), \ldots, (q_{100i}, q_{100v})\}$. The cubic spline function for array i now represents the normalized expression indices. That is, the mapping, from raw to normalized expression indices is given by $x_{ij} \rightarrow s_i(x_{ij})$.

Optimization of preprocessing steps

The material presented below relies both on the topics presented above, and on the topics concerned with statistical testing discussed in Sec. 3.3. Instead of postponing the material presented here to chapter 3, I will simply assume in the following that Sec. 3.3 has been read.

If it were known beforehand which genes would be differentially regulated in an experiment, one could evaluate different combinations of preprocessing steps by their ability to reveal those genes upon subsequent statistical testing. That is, keeping the statistical test the same, by determining the combination of preprocessing steps that allow the largest percentage of the total number of differentially regulated genes to be detected. Of course this should not be at the expense of also detecting a lot of false regulations. What is needed is the combination of preprocessing steps for which the statistical test is maximally

sensitive (TP/TP + FN) and specific (TN/TN + FP). In reality, such a situation is only achieved if the samples used for hybridization have been artificially constructed. For example, by creating samples of 3860 cRNAs each, where 1309 of them differed in varying amounts between three control samples and three 'spiked' samples, Choe et al. (2005) showed that for Affymetrix microarrays, the following combination of preprocessing steps allowed detection of approximately 70% of the true positives at FDR < 0.1: subtracting unspecific signal from the perfect match probe intensities and performing an intensity-dependent normalization at the probe set level. However, usually such spiked datasets are not available and it might even be that the data structure resulting from a given experiment requires a different combination of preprocessing steps than what was determined using a spiked dataset for the same technology platform.

One suggestion for a quality measure for real datasets is to calculate the ratio between biological variance (variance between different sample types) and measurement variance (variance between replicates). Thus, the measure is an F-like statistic averaged over all genes. Thygesen and Zwinderman (2004) have shown that it might be best not to use this ratio directly but instead calculate a p-value under the null hypothesis of no biological variance. The distribution of the ratio, needed for such a p-value calculation, can be computed by constructing a number of permuted datasets and calculate the ratio for each of these.

Another approach is to focus on the results of the statistical testing. For example, for each of a number of preprocessing combinations, by plotting the number of genes detected as differentially regulated (TP + FP) against the number of these genes that are false positives (FP), for a range of significance levels α. One can then choose the combination that performs the best by identifying the curve that has the fewest false positives at any significance level compared to the sum TP + FP (He et al., 2003). To use this method one have to correctly estimate the number of false positives. The best way to do this depends on the experimental design. For example, if several replicates are available, one could randomly divide them into two groups and test for differentially regulated genes between these groups. Any genes detected would constitute false positives. The random division into two groups could be done a number of times and the estimate of the number of false positives be calculated as the average number of genes detected as differentially regulated for each division.

2.1.4 Research example: cDNA array (paper I)

The barleyPGRC1 10K cDNA array

To detect mRNA amplified from barley (*Hordeum vulgare* L., cv. Pallas) epidermal single cell samples as described in Sec. 1.1, a cDNA array (the barleyPGRC1 10K cDNA array, abbreviated the "10k-array") was developed. It consists of two nylon membranes spotted with 10 297 unigenes. In the following, the basic manufacture of the array, as well as the hybridization and preprocessing of the array data, are presented.

The barley genome has not been sequenced yet. However, many ESTs have been created. For the 10k-array, a total of 41 600 barley ESTs of more than

100 bp length were selected from the cDNA libraries generated at the Institute of Plant Genetics and Crop Plant Research, Germany (Künne et al., 2000). Specifically, the HO library, which is primarily derived from barley infected with powdery mildew (*Blumeria graminis* (DC) Speer f.sp. *hordei*), and thus particularly useful to our study, were included. All ESTs were clustered into groups sharing significant sequence similarity. The resulting clusters and singleton sequences (sequences that did not cluster) thus represent a non-redundant set of gene-oriented sequences. That is, each cluster and singleton sequence hopefully correspond to a unique gene, and they are therefore denoted unigenes. When comparing the unigenes on the 10k-array to the genes on the Affymetrix Barley Genome Array (Close et al., 2004), 15% of the genes are unique to our array.

For each unigene, one representative clone associated with the longest EST sequence was selected. Each of the selected cDNA clones was PCR amplified and after purification and two-fold concentration by ultrafiltration, resulting in an estimated DNA concentration of 50 ng/mL, mixed 1:1 with spotting solution. The DNA was then robotically spotted in duplicate onto nylon membranes using a 0.2 mm pin tool. Five strokes of the pin tool were applied to each spot resulting in the transfer of approximately 50 nL material.

Hybridization and scanning

We hybridized 12 10k-arrays with amplified and radioactively labelled material from the cell specific samples (see Sec. 1.1); 3 with material from resistant cells, 3 with material from infected cells, 3 with material from non-inoculated control cells, and 3 identical samples for technical replication, made by pooling equal amounts of material from the 3 infected samples. The radiation from each spot was detected by phosphorimaging and the results stored in TIFF files.

Segmentation and correction for overshining

For radioactively labelled material, the more material hybridized, the larger the spot (see Sec. 2.1.3). Therefore, images were partitioned into spot and background regions using fixed circle segmentation (diameter of circle equal to 0.4 mm, and on average 0.3 mm between circles in the horizontal or vertical direction in the subgrid). For each spot we calculated the average of the corresponding pixel intensities. Due to the above-mentioned radiation effects, much of the intensity measured in background regions, and often also some of the intensity in the spots themselves, contains significant contributions from radiation emitted from neighbouring spots. To correct for this we used a straightforward extension of the approach described by Machl et al. (2002). Briefly, on each nylon membrane we compare the differences between intensities in replicated spots to the differences between the sums of intensities in their respective neighbouring spots. This yields an overshining coefficient c, which is the percentage that each spot overshines into neighbouring spots. Since the overshining coefficient is distance dependent we divide it into two coefficients, one for first neighbours (the four neighbouring spots directly above, below or to the sides), denoted c_1, and one for second neighbours (the remaining four neighbouring spots on the diagonals), denoted c_2. We found that on average for all arrays $c_1 = 0.3\%$ and $c_2 = 0.1\%$. Therefore, especially in the case of low intensity spots with high

intensity neighbours, the overshining has a significant contribution and the correction for it is important. By measuring the background intensity close to the edge of each nylon membrane (where the effect from overshining should be minimal) we found no significant location-dependent variation in the background intensity. Consequently we calculated an overall average background intensity on each array from the pixel intensities around the edge of the membrane.

Data preprocessing

Each of the following steps was performed sequentially. (1) For each spot, the background intensity was subtracted from the spot intensity. (2) Spot intensities were corrected for overshining as described above. (3) All intensities were $\log_2$-transformed. This resulted in reasonably constant variability at all intensity levels. (4) The intensity of each probe was calculated as the mean of the intensities of the two replicate spots on each array. (5) No intensity-dependent overall differences between arrays were detected. Therefore, to remove overall differences in intensities between arrays, the intensity of each probe on each nylon membrane was normalized by subtracting the median of the intensities on each membrane. The background subtracted, overshining corrected, normalized, mean-$\log_2$ intensity of a probe is denoted the expression of the gene corresponding to that probe. All raw and processed data are stored in the ArrayExpress database[13], accession number E-MEXP-457.

Division between plant and fungal genes

The barley cDNA libraries used to design the 10k-array, especially the library made from powdery mildew-inoculated barley, may contain some fungal sequences. To make sure that we only included barley genes in our analyses, fungal sequences were detected and removed as follows: For all genes on the array, sequences that showed high similarity (using both BLASTN and TBLASTX, see Sec. 4.1) to fungal sequences (an E-value less than 10^{-20}) over plant sequences (no hits less than 10^{-20}) were removed. Here, the fungal sequences were from the Cogeme database[14] and the plant sequences constitute the rice (*Oryza sativa*[15]) and Arabidopsis[16] genomes combined. Out of the 10 297 unigenes on the 10k-array, 225 were identified as fungal genes in this manner and removed.

2.2 Proteomics

Loosely speaking, more things can be measured and investigated when it comes to the proteome compared to the genome and transcriptome. For example, proteins get phosphorylated and glycosylated — endowing them with a lot more decorative elaboration than is found in DNA. Of course, DNA also has methylation and a few other modifications, but nothing like the complexity of proteins.

[13] `www.ebi.ac.uk/arrayexpress`
[14] `cogeme.ex.ac.uk/sequence.html`
[15] `ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/`
[16] `ftp.arabidopsis.org/home/tair/home/tair/Sequences/blast_datasets/`

The outline of this section largely follows that of the Sec. 2.1. Techniques for sequencing and identification of proteins are discussed in Sec. 2.2.1. Proteins can interact with each other, forming stable or transient complexes. The reason for forming complexes are numerous and include the ability to increase turnover rates of metabolites by having all pathway enzymes closely together, and to avoid crowding in the confined space of the cell by packing the protein machinery (Ellis, 2001). Experimental approaches for identifying protein-protein interactions are discussed in Sec. 2.2.2. Also discussed in Sec. 2.2.2 are several techniques for large-scale measurement of protein amounts, specifically protein microarrays. Finally, Sec. 2.2.3 presents an example of the development of an antigen microarray.

2.2.1 Sequencing and identification

To determine the amino acid sequence of a protein one can use Edman degradation (Shively, 2000). The Edman degradation sequentially removes one residue at a time from the amino end of a peptide. The separated residue is identified using high-pressure liquid chromatography, HPLC. In HPLC, the residue is forced through a column (stationary phase) by a liquid (mobile phase) at high pressure. The time it takes to pass the column identifies each amino acid uniquely. By repeated degradations, a sequence of around 50 amino acids can be determined in one experiment.

In general, it is easier to sequence mRNA than proteins. And having obtained the mRNA sequence one can easily deduce the resulting amino acid sequence. Therefore, instead of sequencing a protein directly in order to identify it, one can often use the following approach: (1) Digest the protein with a a trypsin, thus creating a number of peptides, (2) measure the molecular mass of each peptide, and (3) compare this with translated known mRNA sequences that has been digested virtually in a similar manner (using an algorithm that implements the properties of the trypsin). The most common method for mass determination is mass spectrometry (MS). Here, an ionizer transfers the peptides into gas phase ions and accelerates them towards a mass analyser. The mass analyzer separates the ions according to their mass/charge ratio. After separation, the ions are then detected by an ion detector (Mann et al., 2001). This is only the basic principle of MS, and different types of MS have been developed that can be used for different applications.

2.2.2 Methods for large-scale proteome analysis

Interactions

Interactions between proteins are primarily determined on a large-scale by one of the following two methods: Yeast-two-hybrid screens (Fields and Song, 1989) or purification of tagged protein complexes followed by mass spectrometry (Aebersold and Mann, 2003).

The basic idea of the yeast-two-hybrid system is to connect the transcription of a reporter gene to the successful interaction between two proteins of interest (say, proteins A and B). That is, only if the proteins interact will the reporter gene

be transcribed. The reporter gene is transcribed when a specific transcription factor (Gal4) binds upstream of it. This transcription factor has two parts, a DNA-binding domain (which binds to the upstream sequence) and an activation domain (which is needed for the RNA polymerase to attach and transcribe the reporter gene). These two parts are separated, the DNA-binding domain fused with the gene encoding protein A, and the activation domain fused with the gene encoding protein B. When the fused genes are transcribed, the chimeric protein A, fused with the DNA-binding domain, will bind upstream of the reporter gene. Only if protein A and B interacts will the chimeric protein B, fused with the activation domain, also be bound in the upstream region of the reporter gene, and the reporter gene be transcribed. The experiment can be carried out thousands of times on microassay plates, as described by Uetz et al. (2000). Disadvantages of the method are that there have been reported high false positive rates (proteins are identified as interacting when they in reality do not), and only pairwise interactions are tested.

In the large-scale purification of protein complexes, fusion proteins are used as well. Here, the protein of interest is fused to a tag. This tag is a peptide that can be used to separate the protein from the solution by binding to an antibody which again binds to an immunoglobulin-bound solid support (affinity purification). When purifying the protein through its tag in this manner, proteins interacting with the tagged protein will also be retrieved. The proteins in the purified complexes are separated by gel electrophoresis and identified by mass spectrometry. Such an approach has been carried out for hundreds of complexes, as described by Gavin et al. (2002).

Besides the experimental approaches outlined above, several computational approaches have been developed to detect putative protein-protein interactions. These usually involve comparisons across species, for example, conservation of gene order, gene fusions, or conservation of phylogenetic profiles (Orengo et al., 2004). Also, genes encoding proteins in stable complexes often display similar gene expression profiles (Hahn et al., 2005).

To reduce the error rates in protein-protein interaction datasets and increase their coverage of the proteome, different datasets can be combined. The simplest rules for integration of datasets is to use 'and' or 'or' rules. That is, all datasets must agree on an interaction (and), or, at least one dataset must predict the interaction (or). Obviously, the 'and' rule results in accurate datasets with low coverage, whereas the 'or' rule trades coverage for accuracy. It has recently been shown (Xia et al., 2004) that using machine learning methods to integrate the datasets in general yields superior results compared to the simple 'and' or 'or' rules.

Currently, around 2000 different types of interactions have been identified (being derived from the total set of all interactions measured), and there is probably around 10 000 different types of interactions all in all (Aloy and Russel, 2004). Consequently, much work still has to be done before a comprehensive coverage of interactions is obtained.

Amounts

In parallel to Northern and Southern blots discussed for RNA and DNA in Sec. 2.1.2, Western blots can be used to detect individual proteins. Traditionally, large-scale analysis of protein amounts have been conducted using 2D-gel electrophoresis (Dunn and Gorg, 2001). Here, a protein mixture is loaded onto a polyacrylamide gel and separated in the first dimension by a pH-gradient (by applying an electric field) and in the second dimension by molecular mass. The proteins are stained and appear as spots in the gel. To identify the proteins in each spot, they are cut out and analyzed by mass spectrometry. There are, however, several problems with the method: The proteins in the gel are degraded over time and consequently one cannot take too long identifying them. Also, large variations between gels from similar samples have been observed. Finally, there may be several proteins in the same spot, making both the functional characterization and identification of such spots difficult.

Specifically for antibodies in, for example, serum (blood plasma without clotting factors), amounts can be measured using the Enzyme-Linked Immunosorbent Assay (ELISA) technique (Lai et al., 2005). The steps of ELISA are: (1) Apply and fix a sample of known antigen to a surface. Often, 96 well plates are used, and each well loaded with a different antigen. (2) Each of the wells are then coated with the serum sample. Antibodies in the serum that are specific to any of the immobilized antigens will bind. (3) The plate is washed so that unbound antibody is removed. After this wash, only the antibody-antigen complexes remain attached in the well. (4) Secondary antibodies are added to the wells. These will bind to any antigen-antibody complexes. (5) The secondary antibodies are coupled to a substrate-modifying enzyme, and when applying this substrate, a chromogenic or fluorescent signal is elicited. (6) The amount of antibody in each well can then be quantified using a spectrophotometer. Such a method has for example been used to analyze patterns of 87 autoantibodies in 20 healthy persons and 20 persons with type 1 diabetes mellitus (Quintana et al., 2003). It was shown that these patterns can discriminate between healthy persons and persons with type 1 diabetes mellitus. However, the method does not easily scale to thousands of antibodies.

As mentioned in Sec. 2.1.2, besides DNA microarrays, much effort has been invested in developing protein microarrays: High-throughput devices for measuring amounts of individual proteins. Proteins are more challenging to prepare for microarrays than DNA, due to for example large differences in binding constants between different proteins immobilized on the array. When reviewing the literature on protein microarray technology, topics such as optimal surface chemistry, capture molecule attachment, and protein labelling and detection, get much attention (Zhu and Snyder, 2003). Several types of protein microarrays have been developed. Depending on what is attached to the solid support, examples are: tissue microarrays (Kononen et al., 1998), peptide microarrays (Fodor et al., 1991), antigen microarrays (Robinson et al., 2002; paper III), universal protein microarrays (Ge, 2000) and carbohydrate microarrays (Wang et al., 2002). The preprocessing of protein microarrays follow the same steps as outlined in Sec. 2.1.3. In the next section, the development and preprocessing of an antigen array will be presented. Before this, however, one remark concerning coverage. Taking humans as an example, compared to the around 30 000

genes comprising the genome, it is currently believed that the proteome consists of around one million different proteins (Stetter et al., 2003). This increase is for example due to post-transcriptional and post-translational modifications as briefly mentioned in the introduction to this chapter. In any given type of cell, only a fraction of the entire proteome is expressed. Still, however, the number of different proteins that need to be measured to provide a reasonable coverage of the entire set of expressed proteins is expected to be quite large, and outside the capability of current technologies. A representative example of current technology being the work by Horn et al. (2006). In this study, a protein array consisting of 37 200 recombinant human proteins was developed to profile the autoantibody repertoire of dilated cardiomyopathy patients.

2.2.3 Research example: Antigen array (paper III)

A total of 266 antigens was selected for the microarray. These includeded proteins, synthetic peptides from the sequences of key proteins, nucleotides, and phospholipids. In general, the antigens can be divided into three groups: (1) those that are known to be involved with diabetes, (2) those that in general are connected with autoimmune diseases, and (3) "chance" antigens, not known to be connected with autoimmune diseases but central with respect to the processes they are involved in. The antigens were diluted in phosphate buffered saline (PBS) and placed in 384-well plates at a concentration of 1 μg/μL. From these wells a robotic arrayer with solid spotting pins of 0.2 mm were used to spot the antigens onto super-epoxi microarray substrate slides. Each antigen was spotted in two to eight replicates. The spotted microarrays were stored at 4°C.

Before incubation with test serum, the microarrays were washed with PBS and blocked for 1 h at 37°C with 1% bovine serum albumin (BSA). They were then incubated overnight at 4°C with a 1:5 dilution of the test serum in blocking buffer under a coverslip. Next, the arrays were washed and incubated for 45 min at 37°C with a 1:500 dilution of a goat anti-mouse IgG Cy3-conjugated antibody. The arrays were washed again, spin-dried, and scanned. The results were recorded as TIFF files.

The pixels that comprised each spot in the TIFF files and the local background were identified by histogram segmentation. The intensity of each spot and its local background were calculated as the mean of the corresponding pixel intensities. None of the spots containing antigens showed saturation. Technically faulty spots were identified by visual inspection and removed from the data set. For each spot, the local background intensity was subtracted from the spot intensity. Spots with negative intensities after this correction were removed from the data set. A $\log_2$-transformation of the intensities resulted in reasonably constant variability at all intensity levels. The log intensity of each antigen was calculated as the mean of the log intensities of the replicates on each slide. To remove overall differences in intensities between arrays, the mean log intensity of each antigen on each array was scaled by subtracting the median of the mean log intensities of all antigens on the array. The scaled mean log intensity of an antigen is denoted the reactivity of the antigen. The processed data set consists of a matrix of IgG reactivities, consisting of 266 rows and 38 columns (two samples for each of the 19 mice). Each column contains the reactivities measured

on a given array and each row contains the reactivities measured for a given antigen over all arrays. Additionally, the reactivity for each antigen measured before and after cyclophosphamide treatment in each mouse was combined into a log ratio by subtracting the reactivity before treatment from the reactivity after treatment. This combination yielded a matrix of ratios with 266 rows and 19 columns.

2.3 Metabolomics

Much important knowledge can be gained from profiling the metabolome. Metabolites, being the intermediates of biochemical reactions, link together a web of interactions, pathways, and participants such as DNA, mRNA, proteins, and environmental stimuli. In essence therefore, the metabolome directly reflects the current status of the cell, and is indicative of the effect of differences in transcript or proteins amounts between different sample types. Concerning the network topology of the metabolome, consider the following example from primary metabolism in fungi (Jewett et al., 2006): (1) More than 70% of all metabolites participate in more than just two reactions, (2) Most reactions require more than one substrate and one product, and (3) the average path length from any metabolite to any other metabolite is around 3. The metabolome is thus highly connected and it might be that small pertubations are propagated to, and therefore detectable in, large parts of the network.

However, the point made in Sec. 2.2.2 concerning the current technological inability to measure the entire proteome is even more pertinent in metabolomics. Here, the current inability to comprehensively profile the entire metabolome is related to its chemical complexity, the biological variance due to e.g. the speed of metabolite turnover, and the dynamic range limitations of the instruments used (Sumner et al., 2003). Multiple technologies are used, including optimized versions of mass spectrometry, nuclear magnetic resonance, and laser induced fluorescence. It is expected that several technologies need to be combined to comprehensively measure the metabolome (Sumner et al., 2003). However, when targeting specific parts of the metabolome where metabolites have reasonably similar properties, using GC/TOF MS, the profiling of as many as a 1000 metabolites have been reported (Weckwerth et al., 2004).

For metabolites that are in dynamic equilibria between free and bound states, the problems of measurement are paramount since they have to be performed in vivo in order not to destroy the equilibria. One such central molecule is NADH, the measurement of which is the topic of the following example.

2.3.1 Research example: NADH monitoring (paper IV)

The basic approach developed for measuring free and bound NAD(P)H is as follows. First, measure and estimate the spectra of free and bound NAD(P)H in vitro. Next, use these spectra to decompose in vivo measured spectra of total NAD(P)H into their free and bound components. The spectrum of free NAD(P)H is easily measured. The spectrum of bound NAD(P)H is estimated

from a series of mixtures that contain varying amounts of free and bound NAD(P)H. Below these steps are described in more detail.

Preparation of mitochondria

Mitochondria were isolated from potato tubers (*Solanum tuberosum* L., cv. Bintje) by differential centrifugation followed by Percoll density gradient centrifugation essentially as described by Struglics et al. (1993). In these mitochondria, no NADPH has been detected (Agius et al., 2001). Consequently, although NADH and NADPH have similar fluorescence characteristics, in the following, we assume that only NADH contribute to the spectra. Total NADH and NAD^+ was extracted from mitochondria and quantified by enzymatic assay essentially as described by Ciotti and Kaplan (1957).

Fluorescence measurements

The fluorescence instrument was developed on-site, specifically for investigating NADH turnover in biological samples. Briefly, fluorescence was induced by the third harmonic of a Q-switched Nd:YAG laser at 355 nm. The laser pulse duration was 5 ns with 1-ns jitter and the pulse repetition rate was 10 Hz. Fluorescence emission was detected through the transparent wall of an oxygraph sample cell, perpendicular to the incident light. The emitted light was collected on the entrance slit of a spectrometer and detected by an ICCD camera. The controller/pulse generator unit coordinated the onset of a laser pulse and the delayed data collection, which was triggered by the Q-switch of the laser. The gate width was 20 ns centered at the peak of the laser pulse intensity, thus ensuring equilibrium between the excitation and the emission processes. Such a data collection window allows determination of the "steady-state" data, where the quotation marks indicate a certain approximation of the term, when applied to our method.

Binding of NADH to model proteins

In order to decompose in vivo measured spectra of NADH into their free and bound components, the spectrum of NADH bound to mitochondrial proteins had to be obtained. Since NADH binds to a wide range of known and unknown proteins, this task cannot be accomplished without certain assumptions. To investigate how the spectrum of bound NADH depends on the binding strength we first measured the binding parameters of several selected proteins by isothermal titration calorimetry, ITC (Leavitt and Freire, 2001). Further, the spectral properties of these systems were studied as described above. The model proteins included a mitochondrial malate dehydrogenase (MDH) from pig heart and a non-mitochondrial protein with an NADH-specific binding site: yeast alcohol dehydrogenase (ADH). BSA has not been reported to have a specific binding site for NADH and was used as a model for unspecific binding of NADH to proteins.

The calorimetric measurements were conducted at 25°C. NADH and/or proteins were dissolved in 5 mM HEPES-KOH, pH 7.5. Concentrations of the

proteins and NADH were determined spectrophotometrically (Hirayama et al, 1990).

The protein solution, typically between 2 and 10 mg/mL, was loaded into the sample cell and titrated with a concentrated solution of NADH. The experiments were designed so that the amount of injected ligand at the end of a titration experiment was twice the number of potential binding sites of the protein. A model with one set of independent binding sites was fitted to the resulting data, and association constants estimated from this. They ranged from $10^4\,M^{-1}$ (in the case of unspecific binding to BSA) to $10^6\,M^{-1}$.

Spectra of free and bound NADH

The fluorescence of NADH in the presence of BSA or ADH shows that the environment of the binding site primarily affects the fluorescence intensity of the bound cofactor and not the spectral shift. That is, binding to an unspecific site (such as in BSA) shifts the emission maximum to the same extent as when NADH is bound to a specific site (as in case of ADH). This ensures that the spectra of free and bound NADH are the same in vivo as in vitro, and consequently that the concentration of free NADH is properly estimated when decomposing spectra. The concentration of bound NADH however, can be over- or underestimated, as the actual average binding constant of mitochondrial proteins is below or above the one assumed from the model system.

From the above considerations, we selected yeast ADH as the model protein with an intermediate between specific and unspecific binding properties for NADH. This model system was routinely tested prior to the beginning of the experiments with the mitochondria.

The spectrum of free NADH was measured directly on the solutions of pure NADH at different concentrations. The linearity of the fluorescence signal with the concentration of NADH was ensured by using low concentrations (i.e. 1–10 μM. Using the binding parameters obtained from ITC, the concentrations of free and bound NADH were calculated for the three standard mixtures of NADH and ADH (Table 3.1 and Fig. 3.9A in Sec. 3.4.1). These systems were designed to contain the same amount of total NAD(H) as our biological system, calculated with the assumption that the isolated mitochondria from potato tubers contain 1–2 nmol NAD(H)/mg protein (Agius et al., 2001). Since we used 1–2 mg protein per mL in the assay, the maximum concentration of NADH in the assay would be 1–4 μM. From the spectra of the standard mixtures the spectrum of bound NADH could be calculated (Fig. 3.9B). This calculation is the subject of Sec. 3.4.1.

2.4 Integration of the omics

Studies that integrate high-throughput data from several omes are appearing more and more frequently. Expression profiles of cell-cycle genes have been combined with protein-protein interaction data to show how functional protein complexes are dynamically assembled and reassembled during the life-cycle of the yeast cell (de Lichtenberg et al., 2005). Gene expression (22 810 genes) and

metabolome (1800 metabolites) data have been combined to study flavonoid biosynthesis in Arabidopsis plants overexpressing the PAP1 gene (Tohge et al., 2005). And Conway and Kinter (2005) identified significantly overexpressed antioxidant and antioxidant-related genes and proteins in connection with the uptake of oxidized low density lipoprotein by macrophages (using Affymetrix microarrays and 2D gel electrophoresis coupled with MS) — just to mention three recent examples. Another obvious application of integrated datasets is for classification purposes of, for example, cancer cell types.

Important for the proper integration of the omes is knowledge about how they relate to each other, and much experimental and theoretical work is still needed in this area. An instructive example of this was presented by Karp (2004) and is the following: The number of biochemically characterized enzymes is around 5000 to 6000. However, only for around 60% of them is the amino acid sequence known. The remaining 40% functionally characterized enzymes cannot be integrated with genomics and transcriptomics data: Genes encoding these enzymes cannot be identified and possible transcriptional control can therefore not be inferred.

Chapter 3

Fitting models to data

Having measured all the compounds constituting each of the omes, as well as their connections, etc., hopefully, the basic experimental knowledge needed to really figure out how it all works together has then been established. That is, the knowledge needed to rationalize the structure and kinetics of the networks in terms of function. To achieve this, however, not only experimental knowledge is needed. Also, theoretical tools that can structure and explain the data, need to be developed and applied (Tilstone, 2003).

This chapter presents examples of fitting models to data from all four papers included in this thesis. As such, the chapter is not meant to provide a comprehensive review of building models and fitting them to data. Rather, the idea is to provide the theoretical background needed for the models built and fitted in the papers. Still, the sections progress from simple to more advanced examples with respect to fitting. In Secs. 3.1 and 3.2, a quantitative method for fitting, the least squares method, is presented. This introduces the basic notation and prepares for the following sections. A simple application of least squares is given in Sec. 3.3. The fits presented in this section, however, only form the starting point for the more general subject of testing for differential expression of compounds between conditions. These methods are used in papers I and III and are also treated in Sec. 3.3. In Sec. 3.4, a geometric interpretation of least squares fitting is presented as preparation for the use of spectral decomposition in paper IV. In general, when fitting models that are linear in the parameters, solutions can be calculated analytically. However, for nonlinear models, e.g. differential equations without closed form solutions, numerical methods must be used. This subject is presented in Sec. 3.5 and applied in papers II and IV. Finally, Sec. 3.6 discusses stochastic models and how they can be used to study cell motility (paper II).

3.1 Maximum likelihood estimation

In essence, the task of fitting a model to data can be formulated as follows: Let the data be a collection of N events corresponding to the measurement of an independent variable x_i and a dependent variable y_i, where $i = 1, \ldots, N$. The

model is some function f that relates the measured independent and dependent variables given a set of model parameters a_j, $j = 1, \ldots, M$, that is

$$f \equiv f(x_i, a_1, \ldots, a_M)$$

When fitting, one wants to adjust the parameters in the model in such a way that each measured y_i is as close as possible to the value predicted by the model at that point $f(x_i, a_1, \ldots, a_M)$. The formulation of a sensible measure of "as close as possible" is the subject of this and the next section.

Often, we can assign a probability to each data point. That is, we might know the probability $\Delta P_i = p_i \Delta y$ that, for a given x_i, any one measurement will give an answer between y_i and $y_i + \Delta y$. Here, p_i is the probability density. Using this, the task of fitting can now be stated as: Given a model with a particular set of parameters, what is the probability that the measured data could have occurred? The parameter values should then be chosen so that this probability is maximized. This is the principle of Maximum Likelihood Estimation.

3.2 Least squares fitting

When a measurement y_i is influenced by many small random and independent perturbations, the probability distribution for this measurement will almost always follow the normal distribution. The reason is that the probability distribution of the sum of a lot of a small random numbers almost always converges to the normal distribution, as proved in the central limit theorem. When using the normal distribution for maximum likelihood estimation, it is usually referred to as least squares fitting. The following calculation shows the reason for this.

Assuming that measurement errors are independent and normally distributed, the probability P that the N events (x_i, y_i) could have come from a model which predicts the values $(x_i, f(x_i, a_1, \ldots, a_M))$, is given by

$$\begin{aligned} P &= \prod_{i=1}^{N} p_i \Delta y \\ &= \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{y_i - f(x_i, a_1, \ldots, a_M)}{\sigma_i} \right)^2 \right] \Delta y \end{aligned} \tag{3.1}$$

where σ_i is the standard deviation of the measurement y_i. Since N, σ_i, and Δy does not depend on the parameters a_j, maximizing (3.1) is the same as minimizing the sum

$$\chi^2 \equiv \sum_{i=1}^{N} \left(\frac{y_i - f(x_i, a_1, \ldots, a_M)}{\sigma_i} \right)^2 \tag{3.2}$$

called the "chi-square". Differentiating (3.2) with respect to the parameters a_j gives a set of equations which must hold at the chi-square minimum.

To summarize: The minimization of the χ^2 represents the quantitative method used to reach the optimal fit of a model to data. Since the χ^2 is the sum of the squared difference between the observed measurements and predicted values,

weighted by the standard deviations; and it is this value we want to minimize, the method is also called least squares fitting. The rationale being that by minimizing this quantity specifically, one maximizes the probability that the model with the determined parameters could yield the measurements.

By error propagation, the variance in each of the parameters can be calculated as

$$\sigma^2(a_i) = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a_i}{\partial y_i} \right)^2$$

which put error bars on the parameter estimates. The overall fit is in general deemed reasonable if χ^2 divided by the degrees of freedom, $\chi^2/(N-M)$, is of order one or less. The reason for this being that if the model is correct, the expected value for χ^2 is equal to $N-M$. Therefore, even with just one set of measurements, if the model is the correct one, it would be highly unlikely that χ^2 is much larger than $N-M$. A more quantitative notion of this is to compare χ^2 to the chi-square distribution for $N-M$ degrees of freedom, and calculate the probability of obtaining a value $\geq \chi^2$ by chance (Bevington and Robinson, 2004).

3.3 Simple models for high-throughput experiments

In high throughput experiments, thousands of genes, proteins, or metabolites are measured under a number of different conditions. For simplicity, gene expression will be used as an example in the following sections, but the methods described are of course generic to all kinds of high-throughput experiments.

Consider a single gene whose expression has been measured under t different conditions. One usually chooses a very simple model to describe the gene's expression under condition k (where $k = 1, \ldots, t$) namely that it is expressed at a given level a_k. Keeping to the notation introduced in Sec. 3.1, for each condition there is one such model, written as

$$f_k = a_k, \quad k = 1, \ldots, t$$

where σ_k is the standard deviation of the measurements. For each condition the gene's expression, y_{ki}, has been measured n_k times (so $i = 1, \ldots, n_k$), and consequently, the χ^2 for each model can be written

$$\chi_k^2 \equiv \sum_{i=1}^{n_k} \left(\frac{y_{ki} - a_k}{\sigma_k} \right)^2$$

Taking the derivative of this with respect to a_k and setting equal to zero gives

$$a_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{ki}$$

So, for each condition, the least squares fit is simply the average of the measured gene expressions.

Having determined the average expression of the gene under each condition is just the beginning. What one is usually interested in is whether the gene is

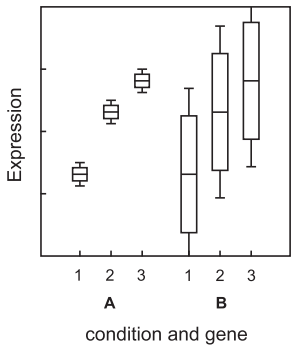

Figure 3.1: Expression values for two genes, A and B, measured at three conditions each. The expression of each gene at each condition is measured a number of times, thus allowing calculation of mean values and standard deviations. These are illustrated with boxplots (the box indicates the standard deviation, the whiskers indicate all measurements).

differentially expressed between conditions. A straightforward approach to this would be simply to use the standard deviations σ_k as error bars on each average expression a_k and inspect whether they overlap or not. Figure 3.1 illustrates this for two genes measured under three conditions each. They have the same mean values for each condition but since the standard deviations for gene A are much smaller than for gene B, it appears more certain that gene A is differentially regulated than gene B. The next section puts this line of thought on more solid statistical ground.

3.3.1 Testing hypotheses, the F-statistic

A gene is either differentially expressed between two or more conditions, or it is not. This can be formulated as the two mutually exclusive hypotheses

$$\begin{aligned} H_0 &: a_1 = a_2 = \cdots = a_t \quad \text{(the average gene expressions are equal)} \\ H_1 &: \text{at least one equality is not satisfied} \end{aligned}$$

where H_0 is the null hypothesis, representing the status quo where the gene is not differentially regulated and H_1 is the alternative (or research) hypothesis where the gene under at least one condition has an expression different from the rest. When testing these hypotheses the basic idea is to calculate a test statistic whose probability distribution can be specified. Using this we can then calculate the probability of incorrectly rejecting the null hypothesis, the p-value. The lower the p-value, the less likely it is that the gene is not differentially regulated.

Since the errors on the measured gene expressions should reasonably be the same no matter the condition, it is in the following assumed that the variances are the same under each condition, that is $\sigma_k^2 = \sigma^2$. Furthermore, to keep things

simple, only the balanced case where $n_k = n$ is considered. The intuition readily extends to the unbalanced case, which just involves lengthier calculations.

Remembering the consideration of the standard deviations of genes A and B in the last section (Fig. 3.1), a straightforward approach for testing the hypotheses is to calculate two independent estimates of the variance σ^2, one of them assuming that the null hypothesis is true. The ratio of these variances could then be compared to the F-distribution, which is the distribution of the ratio of two estimates of the same variance of a normal distribution[1]. The longer out in the tail that the ratio of variances (the F-statistic) is, the less likely it is that the variances are equal.

One estimate of the variance σ^2 can be derived by first calculating the variance for each condition

$$s_k^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ki} - a_k)^2$$

and then take the average of these variances

$$s_{\text{within}}^2 = \frac{1}{t}\sum_{k=1}^{t} s_k^2$$

This estimate has $t(n-1)$ degrees of freedom. Since it is calculated by looking at the measurements within each condition, the subscript "within" was chosen for this estimate.

Another estimate of the variance can be calculated if the null hypothesis is assumed to be true. It is simple to show that the sample mean computed from a random sample of size t from a population with mean μ and variance σ^2 is a random variable with mean μ and variance σ^2/t (Bevington and Robinson, 2004). In the present case we have t conditions with means a_k and variances σ^2. Consequently the variance of the mean expressions is given by σ^2/t. An estimate of this variance is calculated using the mean expressions as observations

$$s_{\text{between}}^2 = \sum_{k=1}^{t}(a_k - a)^2/(t-1)$$

where a is the average of the mean expressions. It follows that ts_{between}^2 is an estimate of σ^2 with $(t-1)$ degrees of freedom. The suffix indicates that it is derived from the variance between the mean expressions of each condition.

The ratio of the two variance estimates

$$\frac{ts_{\text{between}}^2}{s_{\text{within}}^2}$$

is the F-statistic. The larger this ratio, the further out on the right tail of the F-distribution is the statistic, and the less likely it is that the null hypothesis is true. To summarize backwards, the larger the spread is between at least one of the average gene expressions and the rest (s_{between}^2), compared with the spread

[1] Consequently, the F distribution is uniquely identified by the degrees of freedom (number of samples minus one) of each variance estimate.

of values within each condition (s^2_{within}), the less likely it is that the gene is expressed to the same level under all conditions.

When comparing only two conditions, testing based on the F-statistic is the same as performing a two-tailed t-test. However, the F-statistic can be applied to quite complex experimental designs. Consider for example an experiment measured under four different conditions. The experiment is designed to test the effect of two factors, treatment and time, with two levels each: no treatment and treatment, 2 h and 24 h. It is a full factorial design, that is, the four different conditions exactly cover all factor level combinations. One could then proceed with testing the effect of each factor in isolation (main effect) and in combination (interaction effect) by calculating F-statistics for each. Used in this way the analysis is called Analysis of Variance (ANOVA), in the example used here, specifically two-way ANOVA.

When each factor has more than two levels, a problem with this approach is that one does not know which of the levels are different from the rest. A way to remedy this is to use post-hoc comparisons such as for example Fishers least significant difference procedure, Duncan's multiple-range test, or just pairwise t-tests with correction for multiple testing (Freund and Wilson, 2003). However in general, preplanned comparisons have more power. Consequently, if it is known which comparisons (contrasts) are of interest, one may define these to begin with, and then test each one using the F-statistic. An example of such a contrast, using the two-factor experiment mentioned above, could be: Which genes respond differently to treatment at 2 h and 24 h?

When using the F-statistic to test for differential regulation in gene expression experiments, several issues need to be considered. First of all, are the assumptions fulfilled? That is, are replicate measurements normally distributed and have equal variance? And second, in most gene expression measurements, thousands of genes are tested. What cutoff value (significance level) should be chosen to divide genes into those that are, and those that are not, significantly differentially regulated? These are the topics discussed in the next two sections.

3.3.2 Corrections for multiple testing

Consider the situation where m genes are tested. Correspondingly, there will be m p-values. For a given significance level α, genes having a p-value lower than this value are judged as significantly differentially regulated. These genes are termed the positive genes. The genes with p-values larger than α are termed the negative genes. Among the positive genes, some are false positives (FP) in the sense that they are judged as significantly differentially regulated but in reality they are not. The rest of the positives are denoted the true positives (TP). Similarly for the negative genes, some are truly negative (TN), and some are false negatives (FN), that is, they are judged as being not differentially regulated but in reality they are.

When thresholding m genes at significance level α, by the definition of the p-value (the probability of incorrectly rejecting H_0), we thus except the number of FPs to be αm (this of course assumes that the genes are independent, which they in reality are not exactly, because of the networks they interact within). If m is large, αm FPs are usually far too many to include in the group of significantly

differentially regulated genes. A simple correction for this is the Bonferroni correction. Here, the p-value of each gene is multiplied by m (if it then exceeds 1, it is set to 1). This is the same as setting the significance threshold to α/m. With this significance level the number of FPs is $(\alpha/m) \times m = \alpha$. On the other hand, in many situations this often prove to be too restrictive. Also, the number of FNs is quite large when using this correction. More formally, the probability of including at least one FP is called the family wise error rate (FWER), and it is this rate that the Bonferroni correction controls at level α. Several other corrections for multiple testing that control the FWER have been proposed, for example the Sidak, Holm-Bonferroni, Holm-Sidak, Hochberg, and Westfall-Young correction methods (Amaratunga and Cabrera, 2004). They differ by their control over the expected number of FNs, as well as their handling of the not-all-genes-are-independent issue. However, all the correction methods are fairly restrictive, in the sense that the p-values have to be quite small in order to be allowed among the positives.

Often, the main concern is preventing an inordinately large number of false positives from clouding the results. In this case, controlling the expected proportion of FPs among all the positives, termed the false discovery rate (FDR) is what is needed. Benjamini and Hochberg (1995) have proposed a step-up procedure for controlling the FDR. Using this procedure, the p-values are adjusted as follows: (1) Order the m p-values, $p_1 < p_2 < \ldots < p_m$. (2) For each of the p-values in this ordering, adjust p_i to mp_i/i. By allowing only genes with adjusted p-values less than α as the positive genes, the FDR is controlled at level α (Benjamini and Hochberg, 1995). This procedure was later shown to also control the FDR for many types of dependency among the tests (Benjamini and Yekutieli, 2001).

The above adjustment is actually a conservative one. In reality the adjustment is mrp_i/i where r is the fraction of true negative genes, $r = TN/m$. Various approaches for estimating this fraction have been developed (Qian and Huang, 2005).

By controlling the FDR, one usually allows for many more positives than when one controls the FWER. In general, specifically which rate to control and procedure to use depends on what one wants to do with the list of positives generated. For the research in this thesis we found that controlling the "classic" FDR proposed by Benjamini and Hochberg (1995) gave excellent results in terms of biological significance of the list of positive genes. All tests that generate p-values can be controlled with these methods, not just tests based on the F-statistic. For the testing of overrepresented GO terms using Fisher's exact test (see Sec. 4.1), we found that controlling the FWER using the Bonferroni correction was appropriate.

3.3.3 Assumptions

If enough replicates have been made, one can test that the expression measurements of each gene are normally distributed using for example the Kolmogoroff-Smirnoff test (Freund and Wilson, 2003). Also, the assumption of equal variances can be tested using the Hartley F-max test or the Levene test.

For microarray experiments, it is in general accepted that the variation among replicates is roughly proportional to the mean, $\sigma_k^2 \propto a_k$, and thus that a log-

transformation can stabilize the variance. Usually, $\log_2$ is used since two-, four-, eight-, etc.-fold upregulations are then easily identified as the differences 1, 2, 3, etc. between $\log_2$-transformed numbers (and vice versa with negative differences for downregulations). More detailed analysis has shown that for low expression values specifically, it seems that the variance does not depend on the mean. That is, at low expression values, replicate measurements are normally distributed and do not need to be transformed (Durbin et al., 2002). In response to this, the arcsinh function has been proposed as a variance stabilizing transformation that handles both low, medium, and high expression levels correctly (Durbin et al., 2002, Huber et al., 2002).

If the microarray data are very noise and with frequent outliers, or if one does simply not trust, even after transformation, that the data fulfills the assumptions of normality and equal variance, a non-parametric test can be used instead. Two such tests are discussed in the following section.

3.3.4 Non-parametric tests

The non-parametric version of the t-test (two conditions compared) is the Wilcoxon rank-sum test (Freund and Wilson, 2003). There are two groups of intensities we want to compare, and the test proceeds as follows: (1) Replace all intensities with ranks according to their magnitude: 1 for the smallest, 2 for the second smallest, and so on. (2) Next, calculate the sum of ranks in each group. (3) From each sum subtract the theoretical minimum sum (in a group containing one intensity the theoretical minimum sum is 1, in a group containing two intensities it is $1 + 2 = 3$, for three intensities it is $1 + 2 + 3 = 6$, and so on). The two differences are denoted W_1 and W_2 (subscript 1 for the first group and 2 for the second group). (4) Based on W_1 and W_2 and the number of intensities in each group, a p-value can be calculated. Here, the p-value is the probability that the calculated W values could have occurred by chance. Therefore, the lower the p-value, the more significantly different are the two groups of intensities. The idea behind the Wilcoxon rank-sum test is extended to compare more than two conditions in the Kruskal-Wallis test. This test is then the non-parametric equivalent to a one-way ANOVA test.

In general, for the non-parametric tests, robustness and data-structure independence is traded for significance. For a differentially regulated gene where the assumptions are fulfilled, a t-test will usually result in a lower p-value than a rank-sum test.

3.3.5 Modifying the F-statistic when multiple testing

The testing of hundreds or thousands of genes imposes the need for carefully selecting or adjusting p-values as discussed in Sec. 3.3.2. However, besides the need for correcting the p-values due to multiple testing, one can take advantage of the multiple tests performed when calculating the test statistic. For example, all the variances calculated for all genes can be used to estimate the distribution of these. And this variance distribution can be used to identify outlier variances (too large or too small variances) and correct them (by appropriately shrinking

or expanding them). Different empirical Bayes approaches to this have been suggested (Baldi and Long, 2001, Smyth, 2004).

3.3.6 Research example: Detection and validation of genes responding to pathogen interaction (paper I)

Detection

Differential expression between infected and/or resistant samples compared to control samples was assessed using the R software package LIMMA (Gentleman et al., 2004, Smyth, 2004, 2005). This package implements a multivariate statistical method that uses all nine arrays simultaneously to judge statistical significance (as described in Sec. 3.3.5). Briefly, for each gene, empirical Bayes moderated F-statistics were used to test for changes between infected and/or resistant samples compared to control samples. Genes were considered significantly regulated if the p-values, adjusted for multiple testing (Benjamini and Hochberg, 1995), were < 0.05. The false discovery rate (FDR) was thus limited to less than 5%. Using this method, 332 genes were found to be significantly regulated upon powdery mildew attack.

The expression profiles of the 332 candidate genes are shown in a heatmap in Fig. 3.2A, and ordered according to their regulation in a Venn diagram in Fig. 3.2B. The Venn diagram shows that 46 genes were upregulated only in samples from infected cells, 98 genes were upregulated only in resistant cell samples, and 54 genes were upregulated in both infected and resistant cell samples. The numbers of downregulated genes were consistently smaller. Here, 22 genes were downregulated only in infected cells, 73 genes were downregulated only in resistant cells, and 40 genes were downregulated in both infected and resistant cells. No genes were found to be regulated in opposite directions in infected and resistant cells. Thus, our data suggest that in general, if a gene is affected in both types of cells, it is regulated in the same direction. Notice also that on the 10k-array more genes were regulated in resistant cells (265 genes) than in infected cells (162 genes).

Validation

Direct confirmation of all candidate gene expressions in individual resistant and infected cells by e.g. in situ hybridization or single-cell real-time PCR would be extremely laborious, and, crucially, carry implicit dangers of artefactual effects. An alternative is to compare candidate gene expressions with expression data obtained using RNA bulked from dissected epidermal tissue or whole leaves from similar barley/powdery mildew interactions. Such data are not as specific as single-cell data. However, since no genes were found to be regulated in opposite directions in infected and resistant cells, we expect that most genes, if detectable, are regulated in the same direction in single cell and bulk sample extracts. Also, most importantly, use of bulk sample extracts avoids potential artefactual effects on transcript abundance that may be introduced during amplification of mRNA from very small amounts. Thus, comparison with bulk sample extracts would serve not only to validate data for directed regulation of our candidate

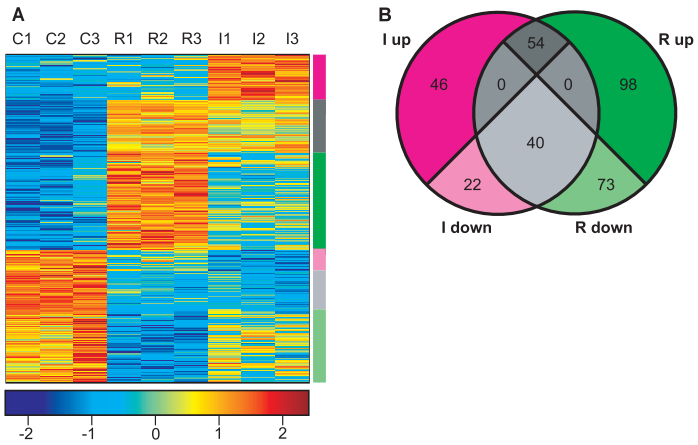

Figure 3.2: The candidate genes and their regulation. (A) Heatmap of the 332 most significantly (FDR < 0.05) differentially expressed genes. The dark red colour signifies high expression whereas the deep blue colour signifies low expression relative to the general expression of each gene. Each column represents a sample and each row represents all expressions for a given candidate gene. The differential expression pattern is clearly seen and can be divided into 6 main patterns of regulation (indicated by the colorbar to the right of the heatmap). (B) Venn diagram grouping the candidate genes according to the 6 regulation types. The numbers in each part of the Venn diagram represents the number of genes with the given type of regulation. C: Control sample, R: Resistant sample and I: Infected sample.

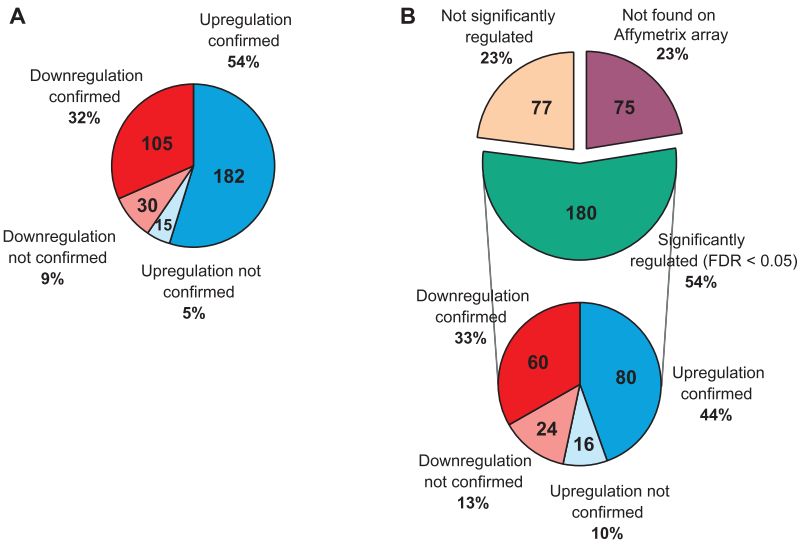

Figure 3.3: (A) Dissected epidermal tissue (10k-array). (B) Whole leaf material (Affymetrix Barley Genome Array).

genes, but would also validate the single-cell analysis approach in general. Consequently, we compared our single-cell data with bulk sample results obtained independently using the 10k-array, the Affymetrix Barley Genome Array, and RT-PCR tests for 10 selected genes.

For comparison with the 10k-array, we considered data obtained using dissected epidermal tissue from an unpublished larger experiment performed at the Gatersleben laboratory, Germany (Axel Himmelbach and Patrick Schweizer, personnel communication). Here, the epidermal tissue was sampled at 24 hours after inoculation (hai) whereas our single cell study extracts were made at 18 hai. Taking the ratio of mean gene expression from inoculated compared to non-inoculated material (four replicates of each) provided regulation data that could be compared to the single-cell data. This analysis confirmed the regulation of 287 out of 332 of our candidate genes, i.e. a confirmation rate of 86% (see Fig. 3.3A for details).

Whole leaf data was available from a time-course study performed by Caldo et al. (2004) using the Affymetrix Barley Genome Array. Using BLASTN we determined the probesets on the Affymetrix array corresponding to the candidate genes on the 10k-array. For these probesets, we then determined which were significantly differentially expressed (using a t-test with p-values corrected for multiple testing FDR < 0.05) between non-inoculated and inoculated whole leaf material extracted at 20 hai. The results are shown in Fig. 3.3B. Out of the 54% of the candidate genes that were significantly regulated, 77% confirmed the regulation found in our single cell study. For these, 90% were also confirmed by the 10k-array data so the confirmed regulations are highly overlapping.

We consider that confirmation rates of 77-86% are sufficiently high to impart

confidence in the regulatory information derived from single cell analyses. This is underlined by the fact that the comparisons were derived: (1) from different laboratories and array platforms and with somewhat different experimental protocols; (2) using RNA bulked from epidermal tissues as well as whole leaves; (3) using slightly different time-points during the interaction; and (4) with different genotypes of barley and powdery mildew fungus. Also, some of the gene regulations that could not be confirmed in bulk extracts could be due to the fact that differential expression of many genes can only be detected in single cell material and not in bulk samples (Gjetting et al., 2004).

To further confirm the observed regulation of the candidate genes, we performed a simple RT-PCR test for 10 of these that are predominantly known to be pathogenesis-related genes. Expressions of these genes were tested in total RNA from powdery mildew-inoculated or uninoculated barley leaf material that was harvested 18 hai. The results confirmed the upregulation of seven and downregulation of two out of three of the genes, i.e. a confirmation rate of 90%.

3.3.7 Clustering and classification

Testing for differential regulation is only one of a plethora of techniques used to analyze different aspects of microarray data. Another much used analysis approach is to investigate relationships between genes or samples. The grouping of genes or samples based on a measure of similarity is known as clustering. Here, the superparamagnetic clustering (SPC) algorithm will be described (Blatt et al., 1996) since that algorithm is used in the example in Sec. 3.3.8 below.

The SPC algorithm can be understood by an analogy to physics: as a parameter T (the temperature) is increased, the system undergoes phase transitions (for example, it melts). In our case, T is increased from 0 (all objects form one cluster) to $T_{\max}$ (each object forms a separate cluster). The breakup of larger clusters into smaller subclusters is governed by the structure of the data: similar objects tend to stay together over a large increase in T, whereas less similar objects break apart more easily. The range of Ts for which a given cluster remains unchanged, denoted by ΔT, is used as a stability measure for the cluster. As the measure of similarity between objects, the Euclidean distance is used for both samples and genes. Because the gene expressions are first row-centered and normalized before being clustered, their squared distance is proportional to $1 - r$, where r is the correlation coefficient. The correlation coefficient captures similarity in shape and the Euclidean distance captures similarity in magnitude.

3.3.8 Research example: Selection of antigen repertoires (paper III)

To ease the following discussion, the experimental protocol is briefly summarized in Fig. 3.4. The numbers in the figure refer to the age (in weeks) of the mice. The black vertical lines at 4 and 9 weeks indicate serum sample collection. The grey vertical lines at 4 and 5 weeks indicate cyclophosphamide injection. The grey box at 6 weeks shows when the CAD-susceptible mice developed diabetes, and the grey box between 11 and 13 weeks shows the time of death of the untreated diabetic mice. In the following, samples taken before cyclophosphamide

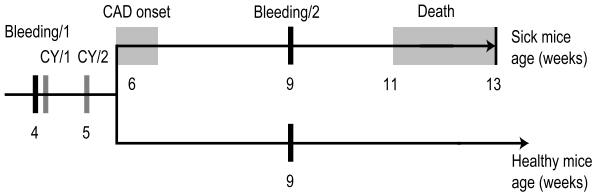

Figure 3.4: The experimental protocol.

injection are indicated by a "B" (before), and samples taken after by an "A" (after). Further, samples from mice that develops diabetes are also indicated by a "S" (sick), and samples from mice that do not develop diabetes are indicated by a "H" (healthy). Thus, there are four sample types: HB, SB, HA, and SA.

Data analysis

To determine subsets of the 266 antigens that would separate the sick and healthy mice, we used the Wilcoxon rank-sum test. This test is non-parametric and therefore robust to outliers. We test one antigen at a time, replacing the reactivities (or ratios) with ranks according to their magnitude: 1 for the smallest, 2 for the second smallest, and so on. The p-values found by using this method were all larger than 0.01; therefore, no single antigen was found to significantly discriminate between the two groups when controlling the FWER (Bonferroni correction) or the FDR (Benjamini and Hochberg, 1995). This finding signifies that the signal produced by any single antigen is unable to separate the sick mice from the healthy. Separability might be achieved, if at all, by using reactivity (or ratio) profiles defined over several antigens.

To capture a collective effect of several antigens, we selected the 27 antigens (10% of the 266 antigens in the study) with the lowest p-values, and investigated how good they were collectively at separating sick from healthy mice, and which of them showed correlated behavior over the samples, by applying two-way SPC. This algorithm gives an unsupervised clustering of the subset of antigens and of the samples, and provides an inherent mechanism for identifying robust and stable clusters in the form of the parameter ΔT. The clusters of samples found by using this method were thus evaluated for their stability ΔT, specificity, and sensitivity. Specificity is the proportion of sick mice in the "sick cluster", sensitivity is the proportion of the sick mice in the sick cluster compared with all of the sick mice in the study.

Statistical significance

To obtain a measure of the significance of the separation between sick and healthy mice by using the method outlined above, we performed the following test: From the group of healthy mice, we picked five of the samples at random, and similarly, for the group of sick mice, we randomly picked four of the samples. These nine samples were labelled as "type 1" The remaining samples were

labelled as "type 1" We hypothesized that there should be no clear separation between these randomized types: we used the Wilcoxon rank-sum test to identify the 27 antigens that differentiated best groups 1 and 2. Next, we clustered the mice in the space of these 27 antigens, and looked for stable, specific, and sensitive clusters, by using SPC. We performed the test 1000 times on different randomized groups and recorded the stability, specificity, and sensitivity of the resulting clusters. The proportion of random clusters manifesting these features to the same or to a better degree than the actual cluster establishes the p-value of the actual cluster.

Selection of informative antigens for pre-CAD mice

Based on the sera taken before cyclophosphamide treatment, the 27 antigens that separated best between the sera of the 10 mice that later resisted the induction of CAD (HB), and of the 9 mice that later developed CAD (SB) (by using the Wilcoxon rank-sum test and taking the 10% with lowest p-values, as described above), were determined. Figure 3.5A shows the two-way SPC of these antigens. The mice susceptible to future CAD induction are denoted by the filled rectangles at the top of the clustering box; the mice resistant to future CAD induction are denoted by the empty rectangles. The 27 antigens are clustered at the rows and are identified by number (refer to Table 1 in paper III). It can be seen that all 9 mice that were found later to be susceptible to CAD could be separated from 8 of the 10 mice that were later found to resist CAD. Before cyclophosphamide, the CAD-susceptible mice manifested relatively elevated IgG reactivity to the top 19 antigens in Fig. 3.5A, whereas the CAD-resistant mice manifested relatively elevated IgG reactivity to the remaining eight antigens. The clustering separation was significant ($p = 0.017$; only 17 of 1000 randomly generated groups showed results comparable with the actual data set). Thus, mice susceptible to CAD could be distinguished by their patterns of IgG serum antibodies from mice resistant to CAD, even before cyclophosphamide was administered to the mice.

Selection of informative antigens for post-CAD mice

We then used the 27 antigens effective in pre-CAD clustering to analyze the patterns of IgG antibodies developing in the diabetic and healthy mice post-CAD. Surprisingly, these 27 antigens failed to discriminate between the two groups of mice; the obvious pre-CAD clusters seen in Fig. 3.5A dispersed when the same antigens were used to cluster the post-CAD sera; compare Fig. 3.5A and B. For this reason, we tested whether other sets of antigens might be more informative post-CAD.

The 27 antigens that best separates the reactivities measured post-CAD (HA versus SA) are shown in Fig. 3.6A. Also, a third set of 27 antigens was generated by performing the Wilcoxon rank-sum test on the ratios by which each antigen changed post-CAD/pre-CAD. The ratios provide information on reactivity changes toward the antigen. These are shown in Fig. 3.6B. It is seen that these two sets of antigens could indeed separate between the healthy and diabetic mice post-CAD: specificity up to 82% and sensitivity up to 100% ($p = 0.065$). Thus, the IgG repertoires of the pre- and post-CAD groups of healthy and sick mice

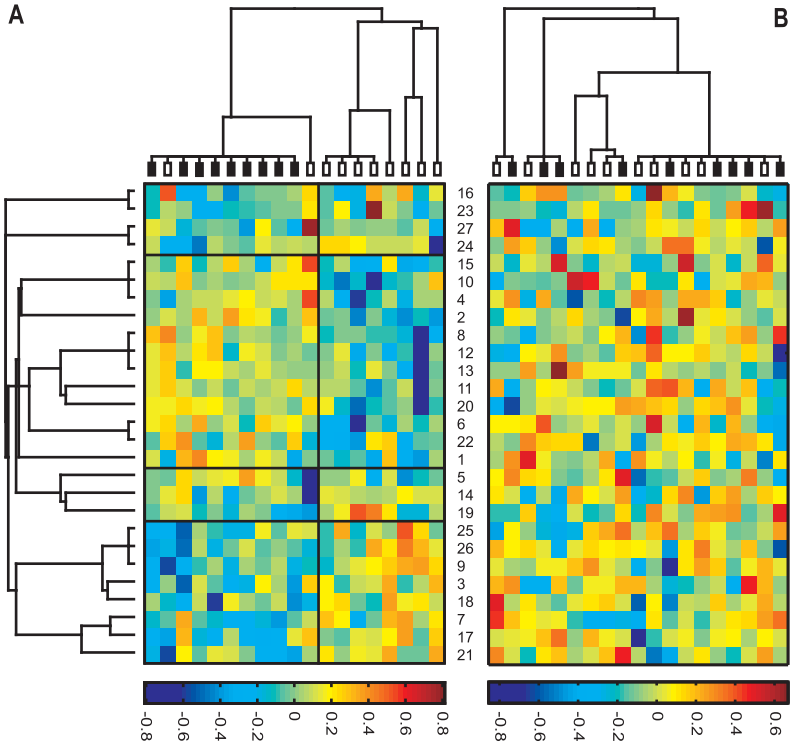

Figure 3.5: Reactivity matrices of 27 antigens that separate diabetic and healthy mice. The rows are antigens and the columns are the mouse sera. Each antigen is identified by the number between the two reactivity matrices (see Table 1, paper III). The Left Panel (A) shows a two-way SPC of the 27 top-ranking antigens from the SB versus HB comparison. The length of a branch connecting to a cluster represents the stability of the cluster. Filled boxes denote mice that later developed CAD, and open boxes denote mice that resisted CAD. The Right Panel (B) shows the SPC of the same antigens, but using the reactivities from the post-CAD samples (SA and HA). Filled boxes denote mice that developed CAD, and open boxes denote healthy mice that resisted CAD.

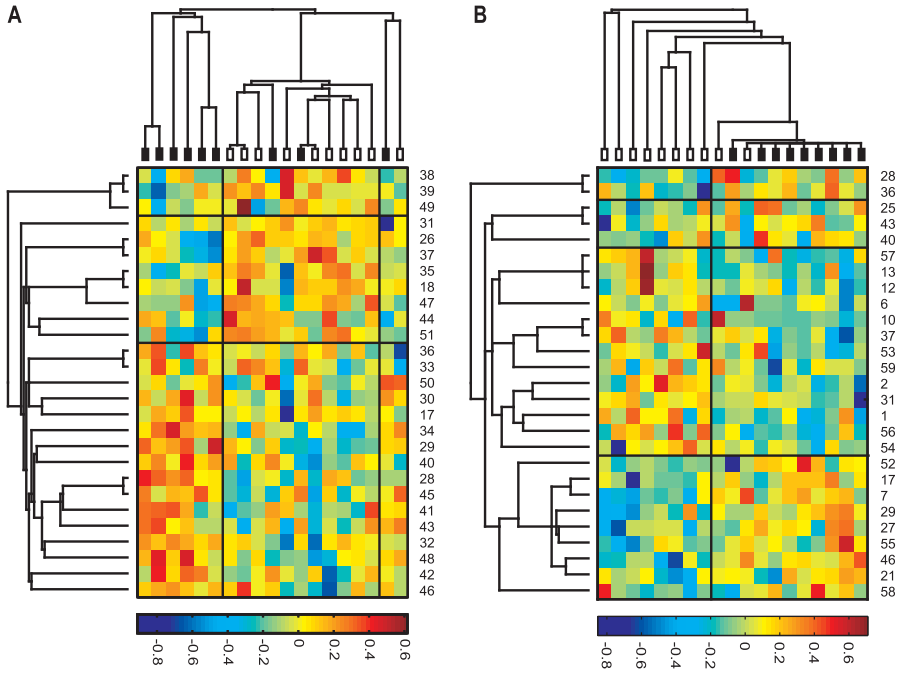

Figure 3.6: (A) Two-way SPC of the 27 antigens (numbered) that best separate SA and HA mice, done applying the Wilcoxon rank-sum test to the reactivities. Filled boxes denote diabetic mice and open boxes denote healthy mice. (B) Two-way SPC of the 27 antigens (numbered) that best separate the sick and healthy mouse samples, done applying the Wilcoxon rank-sum test to the pre-CAD and post-CAD ratios. Filled boxes denote mice susceptible to CAD and open boxes denote mice resistant to CAD.

could be clustered, but the informative patterns of reactivity required modified sets of antigens to develop discriminating patterns. It can be seen that some of the antigens from the set of pre-CAD antigens were also present in the post-CAD set, or in the set of antigens determined from the pre-CAD/post-CAD ratios. For example, three of the pre-CAD antigen reactivities were also prominent post-CAD (antigens 17, 18, and 26). The shared and distinct antigens are shown as a Venn diagram in Fig. 3.7. The group S versus H (ratio difference) can be seen to have generated a set of antigens most shared (dark rectangles) between pre-CAD sera (HB versus SB) and post-CAD sera (HA versus SA).

3.4 Geometric interpretation of least squares fitting

In paper IV, least squares fitting is applied in three situations. The first one being when the binding constants of NADH to several proteins needs to be

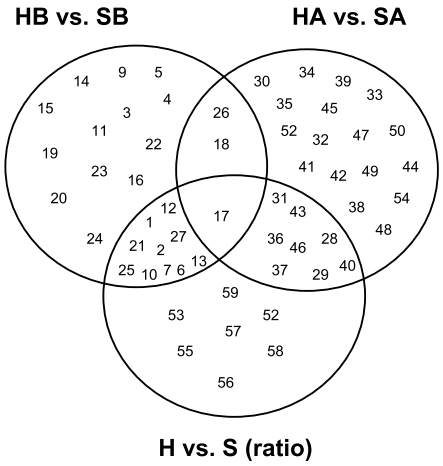

Figure 3.7: Venn diagram showing antigens that were unique to each set or shared by any of the three groups of 27 antigens that best separated the groups of mice.

determined using isothermal titration calorimetry (ITC). This situation will not be discussed here since it is well established (Leavitt and Freire, 2001) and even implemented in a software package that comes with the ITC apparatus. The second being the fitting of the model developed for the homeostasis of free and bound NADH to data (discussed in Sec. 3.5.2). And the third, which is treated in Sec. 3.4.1, being the determination of the spectrum of bound NADH and its subsequent use in estimating the amounts of free and bound NADH in mitochondria. To prepare for this, the geometric interpretation of least squares fitting is introduced below.

As a simple example, take the model

$$f = a_1 x_i + a_2 y_i \qquad i = 1, \ldots, N \tag{3.3}$$

To make the geometric interpretation more interesting there are now two independent variables (x_i, y_i) for each dependent variable z_i (compared to the models introduced in Sec. 3.1 which only had one independent variable). In matrix/vector representation, (3.3) loosely becomes

$$\begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_N \end{pmatrix} = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ \vdots & \vdots \\ x_N & y_N \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

written simply as

$$\boldsymbol{z} = \mathsf{X}\boldsymbol{a} \tag{3.4}$$

where the equality sign is used in the sense that the right side is fitted to the left side. Let the columns of X span the subspace X. The vector $\boldsymbol{z}$ and the

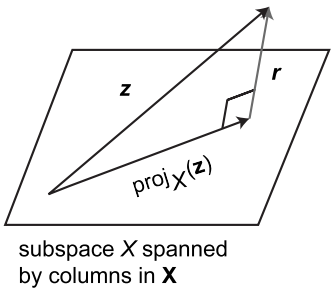

Figure 3.8: Geometric representation of the estimation of $\boldsymbol{a}$.

subspace X is shown in Fig. 3.8. Since all fits of the model in (3.3) to the data in $\boldsymbol{z}$ must lie in the subspace X, it is intuitively clear that the vector in X closest to $\boldsymbol{z}$ is the projection of $\boldsymbol{z}$ onto the subspace X, denoted $\mathrm{proj}_X(\boldsymbol{z})$ (also shown in Fig. 3.8).

That $\mathrm{proj}_X(\boldsymbol{z})$ is the vector in X that is closest, in the least squares sense, to $\boldsymbol{z}$, can be seen from the following argument (adapted from Messer (1994)): Let $\boldsymbol{s}$ be a vector in X. The distance between $\boldsymbol{z}$ and $\boldsymbol{s}$ can be written $\|\boldsymbol{z}-\boldsymbol{s}\|$, where the vertical double bars denote the norm. We need to determine the $\boldsymbol{s}$ that minimizes this distance. First, write $\boldsymbol{z}-\boldsymbol{s} = (\boldsymbol{z}-\mathrm{proj}_X(\boldsymbol{z})) + (\mathrm{proj}_X(\boldsymbol{z})-\boldsymbol{s})$. Since $\mathrm{proj}_X(\boldsymbol{z})$ and $\boldsymbol{s}$ are in the subspace X, so is $\mathrm{proj}_X(\boldsymbol{z})-\boldsymbol{s}$. Since $\boldsymbol{z}-\mathrm{proj}_X(\boldsymbol{z})$ is orthogonal to every vector in X, it is orthogonal to $\mathrm{proj}_X(\boldsymbol{z})-\boldsymbol{s}$. By the Pythagorean Theorem, we can therefore write $\|\boldsymbol{z}-\boldsymbol{s}\|^2 = \|\boldsymbol{z}-\mathrm{proj}_X(\boldsymbol{z})\|^2 + \|\mathrm{proj}_X(\boldsymbol{z})-\boldsymbol{s}\|^2$. The first term in this sum is not affected by our choice of $\boldsymbol{s}$. Inspection of the second term shows that the minimum value of the sum occurs when $\boldsymbol{s} = \mathrm{proj}_X(\boldsymbol{z})$, making this terms zero. That is, by choosing $\boldsymbol{s} = \mathrm{proj}_X(\boldsymbol{z})$, the difference $\|\boldsymbol{z}-\boldsymbol{s}\|^2$, and therefore also the difference $\|\boldsymbol{z}-\boldsymbol{s}\|$, is minimized.

The connection between $\|\boldsymbol{z}-\boldsymbol{s}\|^2$ and χ^2 becomes clear if the former is written explicitly (compare with (3.2))

$$\|\boldsymbol{z}-\boldsymbol{s}\|^2 = \sum_{i=1}^{N}(z_i - s_i)^2$$

Finally, by projecting $\mathrm{proj}_X(\boldsymbol{yz})$ onto each of the columns in X, the parameters a_1 and a_2 are determined.

Using the geometric approach, these projections can be written compactly as follows. First, multiply each side of (3.4) by the transpose of X

$$\mathsf{X}^\mathsf{T}\boldsymbol{z} = \mathsf{X}^\mathsf{T}\mathsf{X}\boldsymbol{a}$$

end then by the inverse of $\mathsf{X}^\mathsf{T}\mathsf{X}$

$$(\mathsf{X}^\mathsf{T}\mathsf{X})^{-1}\mathsf{X}^\mathsf{T}\boldsymbol{z} = (\mathsf{X}^\mathsf{T}\mathsf{X})^{-1}\mathsf{X}^\mathsf{T}\mathsf{X}\boldsymbol{a}$$

By the definition of the inverse of a matrix we can now write

$$\boldsymbol{a} = (\mathsf{X}^\mathsf{T}\mathsf{X})^{-1}\mathsf{X}^\mathsf{T}\boldsymbol{z} \tag{3.5}$$

where $(\mathsf{X}^\mathsf{T}\mathsf{X})^{-1}\mathsf{X}^\mathsf{T}$ is denoted the pseudoinverse of X. Consequently, the pseudoinverse provides a least squares solution to (3.4) (Penrose, 1956). Finally, multiplying with X on both sides of (3.5) gives

$$\mathrm{proj}_X(\boldsymbol{z}) = \mathsf{X}(\mathsf{X}^\mathsf{T}\mathsf{X})^{-1}\mathsf{X}^\mathsf{T}\boldsymbol{z}$$

Having determined the parameters that minimizes χ^2 in this manner, the difference between the model and the data, $\boldsymbol{z} - \mathrm{proj}_X(\boldsymbol{z})$, is denoted the residual, $\boldsymbol{r}$ (shown in Fig. 3.8). The length, $\|\boldsymbol{r}\|$, of which gives an estimate of the quality of the fit.

3.4.1 Research example: Spectral decomposition (paper IV)

The spectrum of free NADH can be directly measured. The spectrum of bound NADH not so. When adding both protein (that binds NADH) and NADH to solution, there will always be a proportion of NADH that is not bound due to the reversibility of the binding process. In the limiting case where large amounts of protein is added to a solution containing only small amounts of NADH, thus making the amount of free NADH very small, the emmision spectrum of free NADH is difficult to detect due to its scarcity. Even if the spectrum is detectable, the system is far from in vivo conditions which makes the relevance of the measured spectra doubtful.

Consequently, in order to determine the spectrum of bound NADH, a model that predicts the combined spectrum of both free and bound NADH is needed. We base the model on the notion that an experimental emission spectrum can be written simply as a linear superposition of the emission spectra of free and bound NADH. Specifically, spectra are nonnegative functions of frequency on the experimentally accessible frequency interval. Squared-integrable real functions on this interval form a vector space. The emission spectra of free and bound NAD(P)H are two linearly independent vectors in this space. The experimental emission spectrum is a third vector in the same space and lies in the two-dimensional subspace spanned by the first two vectors if the experimental emission spectrum originates in free and bound NADH. In practice, we represented the spectra by their values at 1100 frequency values. Then, a vector in the abstract function space is represented by an ordinary vector with 1100 real elements, and decomposition becomes matrix algebra: Emission spectra of the species free NADH and NADH bound to ADH (arranged in an 1100×2 matrix B) were calculated from emission spectra of n standard mixtures of NADH and ADH (in a $1100 \times n$ matrix S) using the $2 \times n$ matrix C of the NADH species concentrations in the standard mixtures as follows. Ideally,

$$\mathsf{S} = \mathsf{BC}$$

Which rearranges to

$$\mathsf{B} = \mathsf{S}\mathsf{C}^\mathsf{T}(\mathsf{C}\mathsf{C}^\mathsf{T})^{-1} \tag{3.6}$$

The standard mixtures used for this calculation are shown in Table 3.1, and their fluorescence spectra in Fig. 3.9A. Specifically, the mixtures 6, 7, and 8 were used in the calculation presented above.

Using the spectra for free and bound NADH thus derived and stored in B (see Fig. 3.9B), the decomposition of experimental emission spectra from solutions

Table 3.1: Calculated concentrations of different NADH species in calibration runs. Concentrations of the stock solutions were determined spectrophotometrically. Measurements were done in a stirred open Rank oxygraph vessel at a solution volume of 0.75 ml.

Mixture	[NADH] (μM)	[ENADH] (μM)	[E] (μM)
1	0	-	-
2	2.22	-	-
3	4.40	-	-
4	6.54	-	-
5	0	0	16.28
6	1.27	0.95	15.20
7	2.58	1.82	14.20
8	3.92	2.62	13.27

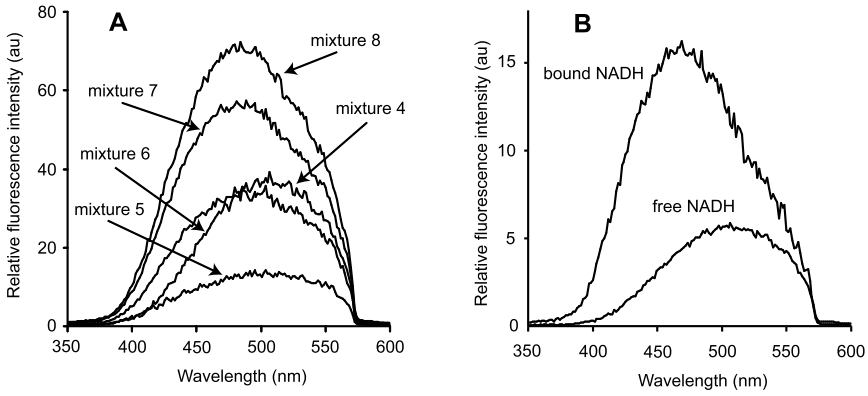

Figure 3.9: Decomposition of standard spectra into free and bound NADH. (A) Pseudo steady-state emission spectra of mixtures of NADH and ADH in 5 mM HEPES pH 7.5 (the exact compositions are given in Table 3.1). (B) The base spectra of free and bound NADH used in the further decomposition of fluorescence spectra of potato tuber mitochondria.

containing isolated mitochondria proceeds as follows. Let the matrix Z contain a series of m experimental emission spectra ($1100 \times m$). The concentrations of free and bound NADH in each experiment in this series (represented by the $2 \times m$ matrix D) multiplied by the emission spectra of free and bound NADH, B, should yield Z.

$$\mathsf{Z} = \mathsf{BD}$$

In a manner similar to above, by taking the pseudoinverse of B, we get

$$\mathsf{D} = (\mathsf{B}^\mathsf{T}\mathsf{B})^{-1}\mathsf{B}^\mathsf{T}\mathsf{Z} \tag{3.7}$$

This decomposition of experimental spectra in Z after the two spectra in B is equivalent with least squares fitting of a linear superposition of the two spectra in B to each of the spectra in Z, as discussed in Sec. 3.4. The parameters fitted are the elements of D, and their resulting values are those given in the identity for D above. The quality of the fit can be evaluated by inspection of the residual spectra arranged in the $1100 \times m$ matrix R as follows

$$\mathsf{R} = \mathsf{Z} - \mathsf{BD}$$

In Fig. 3.10A are shown four experimental spectra (triangles) along with least squares fit of linear superpositions of free and bound NADH (lines). In Fig. 3.10B are shown the residuals for each fit. Notice that the residuals have the same size, no matter the spectrum. This indicates the presence of an additional component which is not dependent on the amounts of free and bound NADH. This could for example be the mitochondria themselves, or simply scattered light from the excitation source.

3.5 Nonlinear models

Nonlinear models are models that result in equations that are not linear in the parameters. For such models, the minimization of χ^2 does not yield a set of coupled equations that are linear in the parameters, and consequently, an analytical solution usually does not exist. Therefore, approximation methods that searches the χ^2 hypersurface to find sets of parameters at local minima are used instead. The generality of such approximation methods also allow models defined by differential equations which have no analytical solution to be fitted to data in this manner. In this case, the fitting algorithm is coupled to the numerical integration of the differential equations.

For interaction networks where the rate equations are known, the network can be modelled by differential equations. Such a network could for example be a metabolic pathway where the workings of each enzyme is specified by a rate equation. The differential equations of such models can not usually be solved analytically, and if they can, the solution is often nonlinear. Another aspect of such models is that usually, the number of parameters that need to be fitted are quite large. For example, for modeling the turnover of NADH in plant mitochondria, Hagedorn et al. (2004) used a network of 7 enzymes and proteins interacting with 14 metabolites. For this relatively simple network, there were 34 parameters that needed to be constrained, specified, or fitted (enzyme

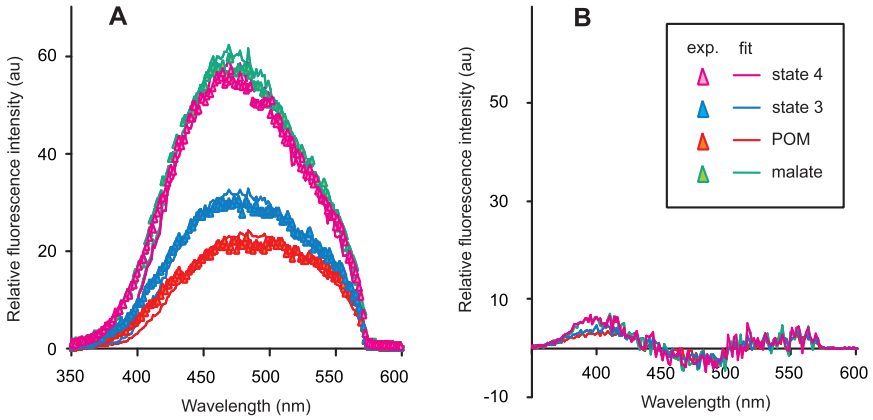

Figure 3.10: (A) Fluorescence spectra from selected time points. Triangles are experimental data, lines are fits. (B) Residuals for the fits shown in A. POM: Mitochondria without substrate or ADP added. Malate: Mitochondria with malate added. State 3: Mitochondria with substrate and ADP added. State 4: Mitochondria with substrate and ADP added, but where ADP has been converted to ATP (see Sec. 3.5.2)

parameters, total metabolite concentrations, addition times, etc.). For models with that many parameters, an immediate concern is that it can be fitted to anything. Ensemble methods that attempt to formalize how such models should be explored and fitted have been developed (Battogtokh et al., 2002, Brown and Sethna, 2003). Another approach that circumvents part of the problem is to focus on the steady-state solution to the model equations, and then expand that solution to the immediate dynamic neighborhood using a nonlinear dynamics framework (Danø et al., 1999).

Several approximation methods have been developed to search the parameter space, for example the Grid Search Method, the Gradient Search Method, and various expansion methods that linearize the fitting function (Bevington and Robinson, 2004). A method that combines these methods is the Levenberg-Marquardt algorithm (Press et al., 1997), which is described in the next section. In Sec. 3.5.2, an example of a simple steady state model of NADH homeostasis is presented and fitted to data using an implementation of this algorithm.

3.5.1 The Levenberg-Marquardt method

For simplicity, let $\boldsymbol{a}$ be a vector containing the parameters of the model. The χ^2 can then be written

$$\chi^2 \equiv \sum_{i=1}^{N} \left(\frac{c_i - y_i(\boldsymbol{a})}{\sigma_i} \right)^2 \tag{3.8}$$

(compare with (3.2) in Sec. 3.2). Differentiating (3.8) with respect to $\boldsymbol{a}$ gives a set of equations which must hold at the chi-square minimum

$$\frac{\partial \chi^2}{\partial a_k} = -2 \sum_{i=1}^{N} \left(\frac{c_i - y_i(\boldsymbol{a})}{\sigma_i^2} \right) \left(\frac{\partial y_i(\boldsymbol{a})}{\partial a_k} \right) = 0, \qquad k = 1, \ldots, M \tag{3.9}$$

By the Levenberg-Marquardt method, the solution to the set of equations (3.9), proceeds iteratively as follows.

The $\boldsymbol{a}$ at the minimum is denoted $\boldsymbol{a}_{\min}$. For an $\boldsymbol{a}$ sufficiently close to $\boldsymbol{a}_{\min}$, denoted $\boldsymbol{a}_{\rm cur}$, it is expected that the χ^2 function (3.8), is well approximated by the quadratic form

$$\begin{aligned} \chi^2(\boldsymbol{a}_{\min}) &\approx \chi^2(\boldsymbol{a}_{\rm cur}) + \sum_i \frac{\partial(\chi^2)}{\partial a_i} a_{diff,i} + \frac{1}{2} \sum_{i,j} \frac{\partial^2(\chi^2)}{\partial a_i \partial a_j} a_{{\rm diff},i} a_{{\rm diff},j} \\ &= \chi^2(\boldsymbol{a}_{\rm cur}) + \boldsymbol{a}_{\rm diff} \cdot \nabla \chi^2(\boldsymbol{a}_{\rm cur}) + \frac{1}{2} \boldsymbol{a}_{\rm diff} \cdot \boldsymbol{D} \cdot \boldsymbol{a}_{\rm diff} \end{aligned} \tag{3.10}$$

where $\boldsymbol{a}_{\rm diff} = \boldsymbol{a}_{\min} - \boldsymbol{a}_{\rm cur}$ and $\boldsymbol{D}$ is an $M \times M$ matrix. Since $\nabla \chi^2(\boldsymbol{a}_{\min}) = 0$, from (3.10)

$$\boldsymbol{a}_{\min} = \boldsymbol{a}_{\rm curr} + \boldsymbol{D}^{-1} \cdot [-\nabla \chi^2(\boldsymbol{a}_{\rm curr})] \tag{3.11}$$

When $\boldsymbol{a}$ is not close to the minimum, or if (3.10) is not a good local approximation, the best thing to do is to take a step down the steepest gradient

$$\boldsymbol{a}_{\rm next} = \boldsymbol{a}_{\rm curr} - \text{constant} \times \nabla \chi^2(\boldsymbol{a}_{\rm curr}) \tag{3.12}$$

where the constant is chosen small enough not to extrapolate the information from $\nabla \chi^2(\boldsymbol{a}_{\rm curr})$ too far away.

In brief, the Levenberg-Marquardt method provides a way to change continuously between (3.11) and (3.12), simply by varying a parameter λ: For $\lambda = 0$, equation (3.11) is used solely, and for $\lambda \gg 1$ equation (3.12) is used solely. For intermediate values of λ, a combination of (3.11) and (3.12) is used.

The method as implemented by Press et al. (1997) then consists of the following steps

(1) Compute $\chi^2(\boldsymbol{a}_{\rm curr})$.

(2) Based on the value for λ, find a new value for $\boldsymbol{a}$, called $\boldsymbol{a}_{\rm next}$.

(3) If $\chi^2(\boldsymbol{a}_{\rm next}) \leq \chi^2(\boldsymbol{a}_{\rm curr})$, decrease λ by a factor of 10. If $\chi^2(\boldsymbol{a}_{\rm next}) > \chi^2(\boldsymbol{a}_{\rm curr})$, increase λ by a factor of 10.

(4) Set $\boldsymbol{a}_{\rm curr}$ to the value of $\boldsymbol{a}_{\rm next}$, and go back to (1).

The fitting stops if no better value for $\boldsymbol{a}$ has been found in 10 consecutive steps, or if a new value for $\boldsymbol{a}$ has only improved $|\chi^2(\boldsymbol{a}_{\rm next}) - \chi^2(\boldsymbol{a}_{\rm curr})|$ by a negligible amount ($< 10^{-3}$), in 5 consecutive steps. Once an acceptable minimum has been found, the $M \times M$ matrix $\boldsymbol{C} = (1/2)\boldsymbol{D}^{-1}$, evaluated at $\boldsymbol{a}_{\min}$, provides an estimate for the covariances of $\boldsymbol{a}$.

The diagonal entries of the covariance matrix $\boldsymbol{C}$ are the squared uncertainties associated with $\boldsymbol{a}$, therefore

$$\sigma(a_j) = \sqrt{C_{jj}}$$

From the off-diagonal elements of $\boldsymbol{C}$, the correlation between elements of $\boldsymbol{a}$ can be determined

$$r(a_{ij}) = \frac{C_{ij}}{\sigma(a_i)\sigma(a_j)}$$

3.5.2 Research example: Modeling NADH homeostasis (paper IV)

When glutamate and malate are added to a solution containing mitochondria, they are transported into the mitochondria and oxidized by enzymes like glutamate dehydrogenase, malic enzyme, and malate dehydrogenase. The oxidation of substrates by these enzymes is coupled to the reduction of NAD^+ to NADH. The two entry-point electron transport chain (ETC) enzymes complex I and ND_{in} then reoxidize NADH to NAD^+, and passes the electrons gained along to the enzymes of the ETC. During this transport the electrons participate in a series of oxidation/reduction reactions before they are finally used by cytochrome c oxidase to reduce O_2 to H_2O. The oxidation/reduction reactions are coupled to the active pumping of protons out of the matrix, and this establishes an electrochemical potential over the inner membrane. This potential is used by ATP synthase to phosphorylate ADP to ATP, be letting 3 to 4 protons flow back into the mitochondrial matrix for each ADP phosphorylated. The production of ATP by the oxidation of substrates, as described here, is called oxidative phosphorylation.

Concerning the turnover rate of NADH during oxidative phosphorylation: When ATP synthase is active and allow protons to flow into the matrix, the membrane potential is smaller, and it is easier for the ETC enzymes to pump protons out again. That is, the ETC enzymes can work faster when ATP synthase is active. Therefore, when ADP is present and ATP synthase therefore is active, the turnover of NADH by complex I and ND_{in} is higher (much higher in fact) than when ATP synthase is not active. When ADP (and substrate) is present, the mitochondria are said to be in state 3, and when ADP is depleted, the mitochondria are in state 4.

In Fig. 3.11 is shown the relative concentrations of total, free, and bound NADH as a function of time. At each time point, these relative concentrations were calculated by decomposing the experimentally measured spectrum onto the spectra of free and bound NADH as discussed in Sec. 3.4.1. As is seen, the amount of free NADH is relatively constant, perhaps displaying a slight increase when ADP is added (state 3) compared to when ADP is depleted (state 4). Since the NADH turnover rate is much higher in state 3 compare to state 4, as discussed above, we expect that there is less NADH present in mitochondria in state 3 compared to state 4. For the amount of bound NADH, this is also the case. However, as seen, the amount of free NADH is constant or even increasing a little in state 3. This indicates that there might be a mechanism which keeps the concentration of free NADH approximately constant.

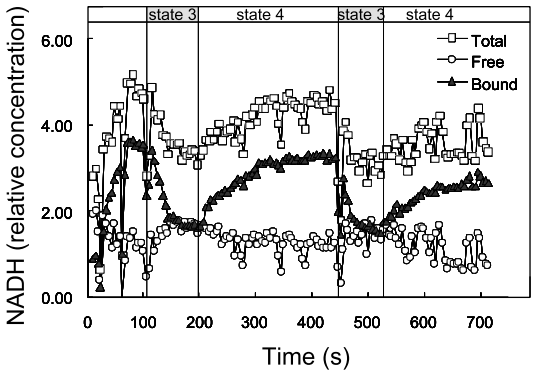

Figure 3.11: Amounts in arbitrary units of free, bound, and total NADH obtained by decomposition of the fluorescence spectra.

We have attempted to explain this phenomenon by modeling the binding of NADH and NAD^+ to mitochondrial proteins in state 3 and 4. Since the amount of experimental data is limited, the model needs to be kept simple in order to avoid having too many unknown parameters, as discussed in the beginning of Sec. 3.5. We proceed as follows.

It is assumed that NADH and NAD^+ bind reversibly to the same pool of binding sites, E, that is

$$\mathrm{E} + \mathrm{NADH} \quad \rightleftharpoons \quad \mathrm{ENADH} \tag{3.13a}$$

$$\mathrm{E} + \mathrm{NAD^+} \quad \rightleftharpoons \quad \mathrm{ENAD^+} \tag{3.13b}$$

where ENADH and $ENAD^+$ denote the bound NADH and bound NAD^+ respectively. In the matrix, NADH is converted to NAD^+ by Complex I and ND_{in}, and NAD^+ is converted back to NADH by (primarily) malic enzyme and malate dehydrogenase. We need not consider the details of these enzymatic reactions if we simply assume that the system has reached a steady state (no change in concentration over time at the time points considered). Consequently, the conversions are written simply as

$$\mathrm{NADH} \quad \rightleftharpoons \quad \mathrm{NAD^+} \tag{3.13c}$$

At steady state, Reactions (3.13a–3.13c) can be written as the following equilibriums

$$K_{\mathrm{NADH}} = \frac{[\mathrm{ENADH}]}{[\mathrm{E}][\mathrm{NADH}]} \tag{3.14a}$$

$$K_{\mathrm{NAD^+}} = \frac{[\mathrm{ENAD^+}]}{[\mathrm{E}][\mathrm{NAD^+}]} \tag{3.14b}$$

$$K_{\mathrm{C}} = \frac{[\mathrm{NAD^+}]}{[\mathrm{NADH}]} \tag{3.14c}$$

where the brackets around the compounds denote equilibrium concentrations, and K_{NADH}, $K_{\mathrm{NAD^+}}$, and K_{C} are the equilibrium constants. Also, reactions

(3.13a–3.13c) imply that the total concentration of binding sites (e) and the total concentration of NAD(H) (n) are constant

$$[\mathrm{E}] + [\mathrm{ENADH}] + [\mathrm{ENAD^+}] = e \tag{3.14d}$$
$$[\mathrm{NADH}] + [\mathrm{NAD^+}] + [\mathrm{ENADH}] + [\mathrm{ENAD^+}] = n \tag{3.14e}$$

Equations (3.14a–3.14e) constitute five equations, with five parameters (K_{NADH}, $K_{\mathrm{NAD^+}}$, K_{C}, e, and n) and five concentrations ([NADH], [NAD$^+$], [E], [ENADH], and [ENAD$^+$]).

Combination of (3.14a–3.14e) yields the equations from which [NADH] (or [E]) can be calculated. First

$$e = [\mathrm{E}]\big(1 + K[\mathrm{NADH}]\big) \tag{3.15}$$
$$n = [\mathrm{NADH}]\big(L + K[\mathrm{E}]\big) \tag{3.16}$$

with

$$K = K_{\mathrm{NADH}} + K_{\mathrm{C}}K_{\mathrm{NAD^+}} \tag{3.17}$$
$$L = 1 + K_{\mathrm{C}} \tag{3.18}$$

The solution for [NADH] can then be written

$$[\mathrm{NADH}] = \frac{-\big(L + K(e-n)\big) + \sqrt{\big(L + K(e-n)\big)^2 + 4nKL}}{2KL}$$

and the other concentrations follow from this.

Next, to obtain values for the parameters, the model needs to be compared to the experimental data. Steady state relative concentrations of free, bound, and total NADH were estimated from Fig. 3.11. Total amounts of NADH and NAD$^+$ were measured by extraction. The estimated relative concentration of total NADH was compared with the measurement of extracted total NADH, and using this comparison, relative concentrations of free and bound NADH were converted to absolute concentrations accordingly. These concentrations are given in Table 3.2.

Assuming that states 3 and 4 are both in steady state, we wish to determine sets of parameters for each state that reproduce the data in Table 3.2 — the most interesting aspect being the slight increase in the amount of free NADH from state 4 to state 3.

There are five parameters: n, e, K_{C}, K_{NADH}, $K_{\mathrm{NAD^+}}$. The extractions show that the total concentration of NAD(H) (n) does not change between states. Likewise, the total concentration of proteins binding NAD(H) (i.e., the total number of binding sites e), is also constant, whereas the binding strength may vary (see below). The parameter K_{C} represents a combination of a number of enzyme reactions, some of which are known to change between states 3 and 4 (e.g., complex I; Hagedorn et al. (2004)). A reasonable value for K_{C}, or an estimate of how it changes between states 3 and 4, is not easily inferred. Consequently, in the following, we do not constrain this parameter between states.

Table 3.2: Total amounts of NADH and NAD^+ were measured by extraction. The values used are specifically those for the same mitochondrial preparation used in Fig. 3.11. The measurement of total NADH was used to convert free, bound, and total NADH as estimated from Fig. 3.11 (state 3, 190 s; state 4, 420 s) from relative concentrations to micromolar in the matrix (1 mg mitochondrial protein is assumed to have 1 μL matrix volume).

Compound	State 3 (μM)	State 4 (μM)
[NADH]	82	62
[ENADH]	82	160
[NADH + ENADH]	160	220
[NAD^+ + $ENAD^+$]	730	640

This leaves us with K_{NADH} and $K_{\mathrm{NAD^+}}$. For these two parameters, the three simplest cases are considered: (1) both K_{NADH} and $K_{\mathrm{NAD^+}}$ are unchanged between states 3 and 4, (2) only K_{NADH} is changed, or (3) only $K_{\mathrm{NAD^+}}$ is changed. For each case, $n = 875\,\mu$M (calculated as the average total NAD(H) in Table 3.2), and we explore a range of values for e (between 200 and 7000 μM). Using these values, the model is least-squares fitted to the data in Table 3.2 using an implementation of the Levenberg-Marquardt algorithm (see Sec. 3.5.1). A fit is considered as reproducing the data in Table 3.2 when $\chi^2 < 1$.

In case 1, keeping both K_{NADH} and $K_{\mathrm{NAD^+}}$ constant, it was not possible to reproduce the data. Thus, it is only possible to model the results in Fig. 3.11 by changing the binding properties of NAD(H). In case 2, choosing an invariable medium binding strength ($K_{\mathrm{NAD^+}} = 400\,\mathrm{M}^{-1}$) similar to that determined previously (Hagedorn et al., 2004, Blinova et al., 2005) and varying K_{NADH}, the data in Table 3.2 were easily reproduced. The values for K_{NADH} determined by this fit are shown in Fig. 3.12A. The ratio, $K_{\mathrm{NADH}}(4)/K_{\mathrm{NADH}}(3)$, lies between 2.5 and 3, as seen in Fig. 3.12B. The absolute value of K_{NADH} depends on the total concentration of binding sites (which is unknown), but in general, a 2.5- to 3-fold reduction of the binding constant in state 3 reproduces the data in Table 3.2. In case 3, the data in Table 3.2 were not reproduced by any reasonable constant value for K_{NADH} (range tested: 10^{-4} to $10^{6}\,\mathrm{M}^{-1}$).

Thus, we conclude that a change in the binding properties of NAD^+, cannot account for the experimental results. However, the data can by reproduced by changing the average binding of NADH to proteins in the matrix and on the inner surface of the inner mitochondrial membrane. Specifically, we estimated that K_{NADH} should decrease 2.5 to 3 times in state 3 compared with state 4.

3.6 Stochastic models

In Secs. 3.6.1 and 3.6.2, the basic theory concerning stochastic modeling of random walks is introduced. This is followed by an example on how such models

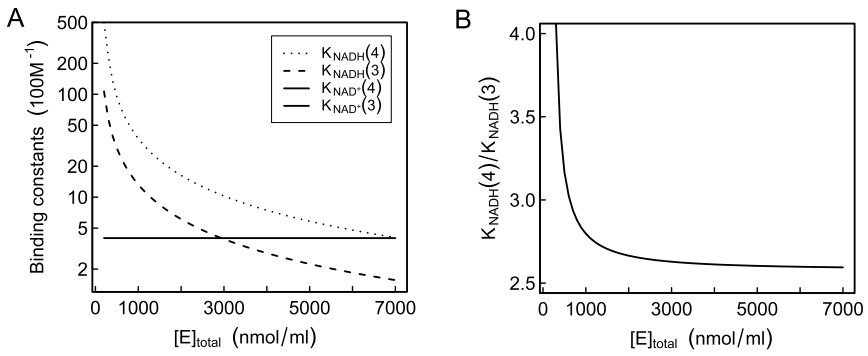

Figure 3.12: Modeling results showing the fitted binding constants. (A) Binding constants as a function of total amount of binding sites as found by fitting. (B) The ratio of fitted binding constants, $K_{\mathrm{NADH}}(4)/K_{\mathrm{NADH}}(3)$, as a function of total amount of binding sites.

can be used for real data in Sec. 3.6.3.

Notation for this section: Expectation (or mean) values are denoted by brackets, for example, the expectation value for $\boldsymbol{\eta}(t)$ is given by $\langle\, \boldsymbol{\eta}(t) \,\rangle$. The covariance matrix of $\boldsymbol{\eta}$ is denoted by $\mathsf{V}(\, \boldsymbol{\eta}(t)\,)$ and the variance by $\sigma^2(\, \boldsymbol{\eta}(t)\,)$. The symbol $\otimes$ is the outer product, for example

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \otimes \begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{bmatrix} 3 & 4 \\ 6 & 8 \end{bmatrix}$$

Matrices are in sans serif, V, and I is the identity matrix. Also, $\delta(t)$ is the Dirac delta function.

3.6.1 The random walk

Noise

The white noise $\boldsymbol{\eta}(t)$ is an uncorrelated stochastic vector with vanishing expectation value, that is

$$\langle\, \boldsymbol{\eta}(t) \otimes \boldsymbol{\eta}(t') \,\rangle \;=\; \mathsf{I}\,\delta(t-t') \qquad \forall\, t, t' \tag{3.19}$$

$$\langle\, \boldsymbol{\eta}(t) \,\rangle \;=\; \mathbf{0} \qquad \forall\, t \tag{3.20}$$

We use the vector notation for the random noise since we will be dealing with random noise in more than one dimension. Equation (3.19) in the definition describes that the random noise is uncorrelated in both time and direction. That is, since the expectation value of two different vector components of the random noise multiplied with each other is always zero, the fluctuation in one direction has no influence on the fluctuation in any other direction. And since the expectation value of the random noise at time t multiplied by the random

noise at time t' is always zero, unless $t = t'$, the noise is uncorrelated in time (the fluctuation a short while ago does not affect the fluctuation now). Basically, the noise has no memory and does not know what each part of it is doing. Equation (3.20) in the definition tells that the expectation value of all components and at all times is equal to zero. That is, the noise fluctuates around zero. This is clearly expected of something we describe as noise.

The reason that the delta function appears in the definition is related to the fundamental inability to measure anything at a point. Rather, we always measure over an interval. The integral of the random noise over a time interval δt, denoted $\boldsymbol{\eta}(t;\delta t)$

$$\boldsymbol{\eta}(t;\delta t) \;=\; \int_t^{t+\delta t} dt'\, \boldsymbol{\eta}(t')$$

is a normally distributed stochastic vector with expectation value equal to zero and variance equal to $N\delta t$, where N is the number of dimensions of $\boldsymbol{\eta}(t)$. This is seen as follows. Since $\boldsymbol{\eta}(t;\delta t)$ is defined as a sum of many random noise terms, the central limit theorem gives that it will be normally distributed. The expectation value is calculated as

$$\begin{aligned}
\langle\, \boldsymbol{\eta}(t;\delta t)\,\rangle &= \left\langle \int_t^{t+\delta t} dt'\, \boldsymbol{\eta}(t') \right\rangle \\
&= \int_t^{t+\delta t} dt'\, \langle\, \boldsymbol{\eta}(t')\,\rangle \\
&= \mathbf{0}
\end{aligned}$$

This is not a surprising result (the expectation value is zero at each time point and the sum of many zeros is still zero). Since the expectation value is zero, the covariance matrix is calculated as

$$\begin{aligned}
\mathsf{V}(\,\boldsymbol{\eta}(t;\delta t)\,) &= \langle\, \boldsymbol{\eta}(t;\delta t) \otimes \boldsymbol{\eta}(t;\delta t)\,\rangle - \mathbf{0}^2 \\
&= \left\langle \int_t^{t+\delta t} dt'\, \boldsymbol{\eta}(t') \otimes \int_t^{t+\delta t} dt''\, \boldsymbol{\eta}(t'') \right\rangle \\
&= \left\langle \int_t^{t+\delta t} dt' \int_t^{t+\delta t} dt''\, \boldsymbol{\eta}(t') \otimes \boldsymbol{\eta}(t'') \right\rangle \\
&= \int_t^{t+\delta t} dt' \int_t^{t+\delta t} dt''\, \langle\, \boldsymbol{\eta}(t') \otimes \boldsymbol{\eta}(t'')\,\rangle \\
&= \int_t^{t+\delta t} dt' \int_t^{t+\delta t} dt''\, \mathsf{I}\, \delta(t'-t'') \\
&= \int_t^{t+\delta t} dt'\, \mathsf{I} \\
&= \mathsf{I}\, \delta t
\end{aligned}$$

The total variance is found by adding the diagonal elements of V

$$\sigma^2(\,\boldsymbol{\eta}(t;\delta t)\,) = \sum_{i=1}^{N} \delta t = N\delta t$$

Notice that the expectation value of $\boldsymbol{\eta}(t;\delta t)$ is still a vector while the variance is a scalar. The more dimensions and the larger the time interval, the larger the variance will be. That is, the more directions the noise has to fluctuate in, and the longer the time interval in which it can do it, the more (measured as variance) will the noise have been away from zero.

The random walk

The random walk performed by a particle with position vector $\boldsymbol{r}(t)$ is described by the differential equation

$$\frac{d\boldsymbol{r}}{dt} \;=\; \rho\,\boldsymbol{\eta}(t) \tag{3.21}$$

where $\boldsymbol{\eta}(t)$ is a random noise as defined above and ρ is a scalar.

Since $\boldsymbol{\eta}$ is a stochastic vector, $\boldsymbol{r}$ is also a stochastic vector. It does therefore not make sense strictly mathematically to take the derivative of $\boldsymbol{r}$. However physically the picture is clear. The particle is constantly changing speed and direction, going hither and dither without any preference in space and time.

To determine where the particle is expected to be at time t we first integrate (3.21) between 0 and t,

$$\boldsymbol{r}(t) - \boldsymbol{r}(0) \;=\; \rho \int_0^t dt'\,\boldsymbol{\eta}(t') \;=\; \rho\,\boldsymbol{\eta}(0;t) \tag{3.22}$$

The properties of $\boldsymbol{\eta}(0;t)$ were derived above. The mean displacement of the particle is given by

$$\langle\, \boldsymbol{r}(t) - \boldsymbol{r}(0) \,\rangle \;=\; \rho\,\langle\, \boldsymbol{\eta}(0;t) \,\rangle \;=\; \mathbf{0}$$

while the mean squared displacement is given by

$$\langle\, (\boldsymbol{r}(t) - \boldsymbol{r}(0))^2 \,\rangle \;=\; \rho\,\langle\, \boldsymbol{\eta}(0;t)^2 \,\rangle \;=\; \rho N t$$

The mean displacement and mean squared displacement of the random walker can be interpreted as follows. Imagine a large number of random walkers all starting in the same point. At time t, each individual walker are in general not at the starting point. But taking the mean of all walker positions at time t gives the starting position, that is the spreading of walkers is symmetrical around the starting position. How far are the walkers from the starting point at time t? The mean squared displacement (which is equal to the variance) gives the answer: at time t, approximately 32% of the walkers are further than $\sqrt{\rho N t}$ away from the starting point.

Discrete random walk

Define

$$\begin{aligned} t_j &\equiv j\delta t \\ \boldsymbol{r}_j &\equiv \boldsymbol{r}(t_j) \\ \boldsymbol{\eta}_j &\equiv \boldsymbol{\eta}(t_j,\delta t)/\sqrt{\delta t} \end{aligned}$$

where $\boldsymbol{\eta}_j$ as defined here is a stochastic vector with zero mean and unit variance in each dimension. Using (3.22) we can now write

$$\boldsymbol{r}_{j+1} = \boldsymbol{r}_j + \rho\sqrt{\delta t}\,\boldsymbol{\eta}_j \tag{3.23}$$

This discretized version of the random walk[2] is easily implemented on a computer.

3.6.2 The persistent random walk

The persistent random walk is described by the following stochastic differential equation (SDE)

$$\frac{d\boldsymbol{v}}{dt} \;=\; -\beta\,\boldsymbol{v}(t) + \sigma\,\boldsymbol{\eta}(t) \tag{3.24}$$

where β and σ are scalars, $\boldsymbol{v}(t)$ is the velocity at time t of the persistent random walker, and $\boldsymbol{\eta}(t)$ is a random noise as defined above. The *persistence time* of a persistent random walk is defined as $P \equiv 1/\beta$.

Discrete persistent random walk

The difference between the current velocity $\boldsymbol{v}(t)$ and the velocity a time δt later, $\boldsymbol{v}(t+\delta t)$, is given by

$$\begin{aligned}
\boldsymbol{v}(t+\delta t) - \boldsymbol{v}(t) &= \int_t^{t+\delta t} dt'\,\frac{d\boldsymbol{v}}{dt} \\
&= \int_t^{t+\delta t} dt'\,\big(-\beta\,\boldsymbol{v}(t') + \sigma\,\boldsymbol{\eta}(t')\big) \\
&= -\beta\,\boldsymbol{v}(t)\,\delta t + \sigma\,\boldsymbol{\eta}(t;\delta t)
\end{aligned} \tag{3.25}$$

where the calculation of the first term in (3.25) is based on the first order approximation that $\boldsymbol{v}(t')$ is essentially constant over the short time interval δt.

Just as above we define

$$\begin{aligned}
t_j &\equiv j\delta t \\
\boldsymbol{x}_j &\equiv \boldsymbol{x}(t_j) \\
\boldsymbol{v}_j &\equiv \boldsymbol{v}(t_j) \\
\boldsymbol{\eta}_j &\equiv \boldsymbol{\eta}(t_j,\delta t)/\sqrt{\delta t}
\end{aligned}$$

From Eqn. 3.25 we can now write

$$\boldsymbol{v}_{j+1} - \boldsymbol{v}_j \;=\; -\beta\,\boldsymbol{v}_j\,\delta t + \sigma\sqrt{\delta t}\,\boldsymbol{\eta}_j$$

which can be rewritten as

$$\boldsymbol{v}_{j+1} = \;(1-\delta t\,\beta)\boldsymbol{v}_j + \sigma\sqrt{\delta t}\,\boldsymbol{\eta}_j \tag{3.26}$$

[2] which is equivalent to with the Wiener process, as used by statisticians (compare with Eq. 2.1 in Higham (2001)).

3.6.3 General models for cell motility

The Ornstein-Uhlenbeck process

The persistent random walk presented in Sec. 3.6.2 has been used to model cell motility in numerous studies, the first time probably by Gail and Boone (1970). When used in this context it is known as the Orstein-Uhlenbeck (OU) process (Ornstein, 1919, Uhlenbeck and Ornstein, 1919), and usually written

$$P\frac{d\boldsymbol{v}}{dt} \;=\; -\boldsymbol{v}(t) + \sqrt{2D}\,\boldsymbol{\eta}(t) \tag{3.27}$$

(compare (3.27) with (3.24) where $\beta = 1/P$ and $\sigma = \sqrt{2D}/P$).

Reading data

As will be shown in the following, the OU process cannot be used to model the motility of HaCaT and NHDF cells. This can be seen by plotting the motility data in various ways that bring forth the basic characteristics of the cells' behavior. Additionally, such plots also point at the kinds of models that do capture the observed behavior. For HaCaT cells, the plots are shown in Fig. 3.13. The individual plots will be presented below, and their agreement or disagreement with the OU model discussed.

Equation (3.27) states that a cell's acceleration $d\boldsymbol{v}/dt$ at any time t is a random vector with expectation value $-\boldsymbol{v}(t)/P$, and with equal RMSD in all directions. This statement can be compared with experimental data in a straightforward and model-independent manner.

Figure 3.13A shows experimentally measured accelerations of HaCaT cells on a substrate of collagen plotted against their instantaneous speed. If these data can be described by (3.27), Fig. 3.13A1 accelerations parallel to the velocity should average locally, as function of the speed v, to $-v/P$, with P a parameter to be fitted. The red data in Fig. 3.13B show that they do. Also, Fig. 3.13A2 accelerations orthogonal to the velocity should average to 0 locally, for all values of v. The green data in Fig. 3.13B show that they do.

Furthermore, the accelerations in Fig. 3.13A should scatter about these averages with the same v-independent RMSD, $\sqrt{2D/\Delta t}/P$, according to the OU model (see also (3.26)). Here $\Delta t = 15\,\mathrm{min}$ is the time lapse between successive position measurements $\boldsymbol{r}_j$ from which velocities and accelerations were calculated; see Sec. 1.4. The blue and magenta data in Fig. 3.13B show the RMSD of each component of the accelerations. Within their error bars, the two RMSDs are identical. But they depend on the speed v, essentially as a first-degree polynomial, in disagreement with the OU model.

Finally, the accelerations in Fig. 3.13A should be normally distributed about their averages and uncorrelated, according to the OU model. Fig. 3.13C shows clearly that they are not normally distributed. Both are more similar to two exponential distributions placed back-to-back. This is not just due to time-lapse recording, i.e., a discretization effect due to $\Delta t = 15\,\mathrm{min}$. One can prove that the OU process gives a purely normal distribution when studied with any finite value for Δt.

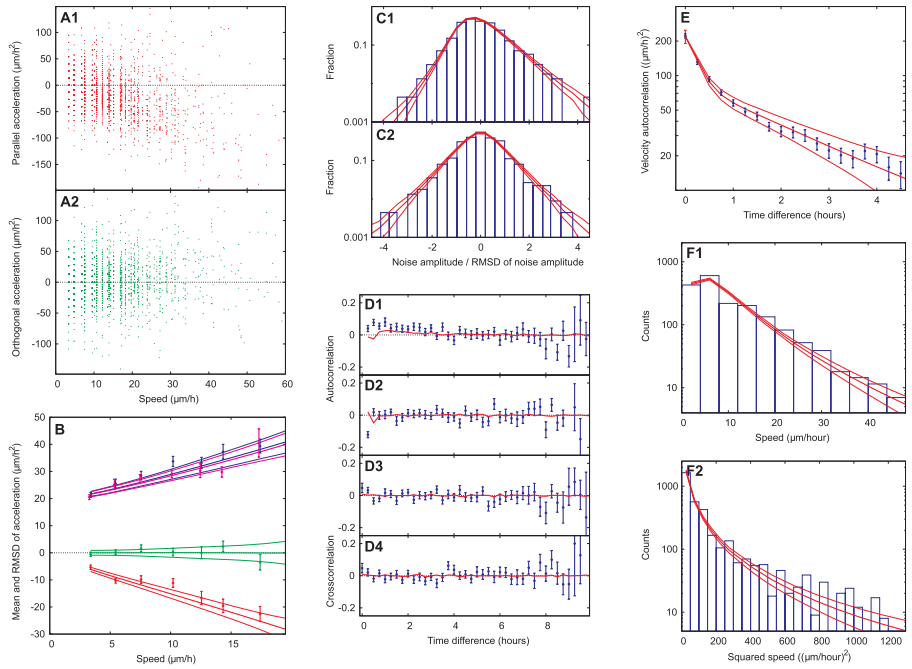

Figure 3.13: Motility of HaCaT cells on collagen surface.
(A) Experimentally measured components of acceleration parallel (A1) and perpendicular (A2) to the cell's velocity, plotted against the cell's speed.
(B) Red and green data points are the mean values of the acceleration's two components shown in panels A1 and A2, respectively, as functions of speed. Blue/magenta data points show RMSD of same quantities, parallel/orthogonal component. Thick curves show theoretical expectation values of same quantities, according to the HaCaT model, fitted to data in panels B, E, and F1 simultaneously. Pairs of thin lines next to thick curves show ± 1 SD from thick curve, according to HaCaT model's prediction of experimental data points' scatter.
(C) Distribution of acceleration about its mean value, in units of distribution's RMSD, averaged over speeds v larger than $0.5\langle v\rangle$. Panels C1/C2 show distribution of component parallel/orthogonal to the velocity. Histograms show experimental results. Thick curves (averages) surrounded by two thin curves (average ± 1 SD) show predictions of the HaCaT model, not a fit to data shown here.
(D) Correlation functions for acceleration's scatter about its mean. Panels D1/D2 are the autocorrelation function of parallel/orthogonal component, normalized to unity at time difference zero. Panels D3/D4 are the cross-correlation function between parallel and orthogonal component, for positive/negative time difference, in units of (RMSD of parallel component)$\times$(RMSD of orthogonal component). Curves are theoretical expectation values.
(E) Data points are experimental velocity autocorrelation function $\phi(t)$. Error bars on data points underestimate true scatter as they were computed from experimental data that are correlated due to persistence of cell motion. The thick line is a guide to the eye connecting theoretical result for $\phi(t)$, computed from 15-min time-lapse position measurements, exactly as in experiment. Two thin lines are ± 1 SD on theoretical result.
(F1) Histogram is distribution of observed speeds v, binned on v-axis, and plotted with lin-log axis. Thick curve is HaCaT model's distribution of speeds. (F2) Histogram shows distribution of observed speeds v, binned on v^2 axis, and plotted with lin-log axis, so a Gaussian velocity distribution $p(\boldsymbol{v})$ would fall on a straight line. Curves are the same as in panel F1.

That the noise is uncorrelated, just as in (3.19), is shown in Fig. 3.13D.

From (3.27), the velocity autocorrelation, $\phi(t)$, can be calculated as

$$\phi(t) \equiv \langle \boldsymbol{v}(t) \cdot \boldsymbol{v}(t) \rangle = \frac{nD}{P} e^{-|t|/P}$$

Fig. 3.13E shows that the velocity autocorrelation function for HaCaT cells is not a simple exponential function, as it is in the OU process. On the contrary, the data are fitted perfectly by a sum of two exponentials.

Fig. 3.13F shows the experimental result for the velocity probability distribution, $p(\boldsymbol{v})$, as a histogram of observed velocities, binned on the v-axis in panel F1, and binned on the v^2 axis in panel F2, both with logarithmic second axis. Isotropy of the surface on which the cells crawled, makes $p(\boldsymbol{v})$ depend only on the speed v, and not on the direction of the velocity. Consequently, if $p(\boldsymbol{v})$ is a Gaussian distribution on the $\boldsymbol{v}$-plane, as in the OU model, its graph is a straight line in panel F2. This is clearly not the case, according to the histogram of observed velocities.

Fig. 3.14 shows results for NHDF cells on TCPS. They move approximately twice as fast, on average, and are also in other ways more dynamic than keratinocytes. Their mean acceleration parallel to their velocity, e.g., decreases somewhat faster than proportional to v (red data points in Fig. 3.14B). Also, the RMSD of the acceleration differs for the two directions: its component orthogonal to the direction of motion is almost constant as in the OU model. But its component parallel to the direction of motion increases with v and doubles its value in the window shown.

Properties of an unknown model

Figure 3.13E shows that the velocity autocorrelation function for HaCaT cells is fitted perfectly by a sum of two exponentials. We consequently assume that

$$\phi(t) = \phi_1 e^{-|t|/P_1} \phi_2 e^{-|t|/P_2} \tag{3.28}$$

and ask which generalization of the OU model might yield this function and the other properties of the data shown in Fig. 3.13. Such a generalization must, like the OU model, describe the rate of change of the cell's velocity. This acceleration cannot depend on a cell's position $\boldsymbol{r}$; nor explicitly on spatial direction or time, because the cells crawl on surfaces that are homogenous and isotropic, and their environment is kept constant. So the acceleration must depend only on the cell's velocity, like in the OU model. We expect some memory in a cell, however, so the acceleration may depend not only on the current velocity, as in the OU model, but on past velocities as well. This dependence must be linear, like in the OU model, to ensure that $\langle d\boldsymbol{v}/dt \rangle_{\boldsymbol{v}} \propto \boldsymbol{v}$ as in the red data in Fig. 3.13B. The noise term seems to have an isotropic amplitude σ, as in the OU model, because blue and magenta data in Fig. 3.13B coincide. But unlike the OU model's amplitude, this amplitude must be speed dependent, $\sigma = \sigma(v)$, according to the same data. The noise itself was uncorrelated to a good approximation, when measured with our 15-min time resolution; see Fig. 3.13D. We consequently model the noise as uncorrelated on all timescales, to keep the model as simple as our data allow.

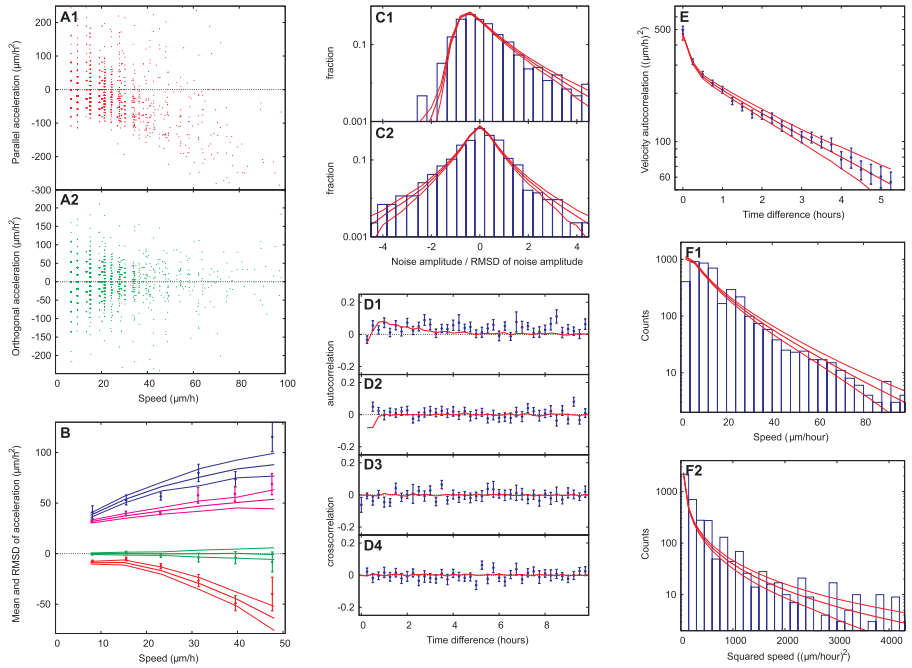

Figure 3.14: Same as Fig. 3.13, but for NHDF cells on surface of TCPS.
(A) and (B) Experimentally measured components of acceleration and their mean (red and green) and RMSD (blue and magenta) values. For details about these panels and the rest in this figure, refer to Fig. 3.13.
(C) Distribution of accelerations. (C1) The asymmetry of the distribution for NHDF cells is more pronounced than for HaCaT cells. The NHDF model's prediction for this distribution and that in panel C2 is shown as red lines. It is not a fit. The model was fitted only to data in panels B, E, and F1.
(D) Correlation functions. (D1) Red curve is the NHDF model's prediction for the autocorrelation function for the acceleration's scatter about its mean in direction parallel to the velocity. Like the experimental data, it is small, but nonvanishing, though the model's noise term is uncorrelated Gaussian.
(E) Velocity autocorrelation function. (F) Distribution of speeds.

This list of properties of the yet unknown model, narrows it down to the integro-differential equation

$$\frac{d\boldsymbol{v}}{dt} \;=\; -K * \boldsymbol{v} + \sigma(v)\,\boldsymbol{\eta}(t) \tag{3.29}$$

where K is a memory kernel yet to be determined, except causality demands that $K(t) = 0$ for $t < 0$: the future must not affect the present. The asterisk denotes convolution, i.e., $(K * \boldsymbol{v})(t) = \int_{-\infty}^{t} K(t-t')\boldsymbol{v}(t')dt'$. K is a scalar function of t, as opposed to a tensor of rank two, because the surface on which the cells move, is isotropic to them (see Sec. 1.4). When this equivalence principle is enforced on the theory we wish to find, mathematics gives that K must transform under spatial rotations as an invariant tensor, i.e., be a scalar. If the isotropy of space is broken, as it is in chemotactic and galvanotactic experiments, K is a tensor of rank two.

HaCaT model

Equations (3.28) and (3.29) have only one solution for K as shown in paper II. Inserted in (3.29), it results in the equation of motion

$$\frac{d\boldsymbol{v}}{dt} \;=\; -\beta\boldsymbol{v}(t) + \alpha^2 \int_{-\infty}^{t} dt' e^{-\gamma(t-t')}\boldsymbol{v}(t') + \sigma\big(v(t)\big)\boldsymbol{\eta}(t) \tag{3.30}$$

Here α, β, and γ are known functions of P_1, P_2, and ϕ_1/ϕ_2, and satisfy $\beta\gamma/\alpha^2$, as they must for velocities to remain finite under the dynamics of (3.30); see Appendix A, paper II. The term in (3.30) containing β represents "loss of memory" of velocity at average rate β^{-1}. The term containing α^2, on the other hand, represents "memory" with characteristic time γ^{-1} and strength α^2/γ.

Inspired by the data in Fig. 3.13B, we choose

$$\sigma(v) \;=\; \sigma_0 + \sigma_1 v \tag{3.31}$$

Equations (3.30) and (3.31) defines our model for HaCaT cells.

A small trick simplifies numerical integration of (3.30) as well as some analytical considerations. Rather than doing the convolution integral in (3.30) at each time step of its integration, we introduce the auxiliary velocity

$$\boldsymbol{V}(t) \;=\; \alpha \int_{-\infty}^{t} dt' e^{-\gamma(t-t')}\boldsymbol{v}(t)$$

and solve a mathematically equivalent problem of two coupled, but ordinary, SDEs,

$$\begin{aligned} \frac{d\boldsymbol{v}}{dt} &= -\beta\boldsymbol{v}(t) + \alpha V(t) + \sigma\big(v(t)\big)\boldsymbol{\eta}(t) \\ \frac{d\boldsymbol{V}}{dt} &= \alpha\boldsymbol{v}(t) - \gamma\boldsymbol{V}(t) \end{aligned}$$

the discretization of which follows along the same lines as in Secs. 3.6.1 and 3.6.2.

The red lines in Fig. 3.13 show the result of fitting the HaCaT model to data (using an implementation of the Levenberg-Marquardt algorithm as discussed in Sec. 3.5.1).

NHDF model

The behavior of NHDF cells, as shown in Fig. 3.14 and briefly discussed above, is captured by two modifications of (3.30): Firstly, we replace the constant β with an increasing function $\beta(v)$. A first-degree polynomial, $\beta(v) = \beta_0 + \beta_1 v$, turns out to do the job, whereas a second degree polynomial does not further improve the agreement between theory and data. Secondly, we replace the scalar function $\sigma(v)$ with the tensor $\bar{\boldsymbol{\sigma}}(\boldsymbol{v}) = \sigma_\parallel(v)\hat{\boldsymbol{v}} \otimes \hat{\boldsymbol{v}} + \sigma_\perp(v)\check{\boldsymbol{v}} \otimes \check{\boldsymbol{v}}$. Here $\hat{\boldsymbol{v}} = \boldsymbol{v}/v$, $\check{\boldsymbol{v}}$ is a unit vector orthogonal to $\boldsymbol{v}$; and $\sigma_\parallel(v)$ and $\sigma_\perp(v)$ are the RMSDs of the random components of the acceleration, parallel and orthogonal to the velocity, respectively. The data in Fig. 3.14 suggest that $\sigma_\parallel(v)$ and $\sigma_\perp(v)$ are different first-degree polynomials in v, but with $\sigma_\parallel(0) = \sigma_\perp(0)$, so $\bar{\boldsymbol{\sigma}}(\boldsymbol{v})$ contains a total of three free parameters to be fitted, where $\sigma(v)$ contained two. This defines our motility model for NHDF cells.

The red lines in Fig. 3.14 show the result of fitting the NHDF model to data.

Same cells on different surfaces

Motility data were collected for NHDF cells on four different surfaces: glass, TCPS, fibrillar collagen, and molecular collagen, and for HaCaT cells on fibrillar collagen and molecular collagen (HaCaT cells did not show measurable motility on glass and TCPS during the first 24 h of the experiment). The results when fitting the models derived above to each of these datasets (see paper II), show that the models are cell-type specific in the sense that they capture the motility patterns on each surface; only the parameters vary. Consequently, these parameter values can be used to characterize the surface as well as the range of responses of the cell.

Chapter 4

Integrating biological knowledge

In this chapter, initiatives to structure and order the body of biological knowledge in ways that are accessible to automated computer queries are presented. In Sec. 4.1, the basic issues of describing gene and protein function and comparing their sequences, are described. Building on this, Sec. 4.2 describes efforts to formalize the descriptions of how genes, proteins, and metabolites relate to one another through networks. Research examples of these issues are given in Secs. 4.3–4.5, all from paper I. Finally, Sec. 4.6 discusses methods to automatically extract knowledge from literature databases — text mining. Such automatic extractions can be used both for increasing the amount of structured knowledge as discussed in the first sections of the chapter, and to retrieve information about the specific results of an experiment.

4.1 Annotation, sequence similarity, and controlled vocabularies

When a gene or protein and its function is discovered through experiments, a brief description of this function is included when the gene or protein sequence is uploaded to a public database (such as GenBank[1]). With the advent of high-throughput sequencing, many putative genes and unigenes have been identified whose functions have not yet been investigated in detail. Such sequences are compared to genes and proteins with known function, and if genes or proteins with significant sequence similarity are found, their annotation are transferred to the new genes. These functions can then be confirmed or disproved later by directed experiments.

The most commonly used algorithm for comparing sequences in this manner is the BLAST (Basic Local Alignment Search Tool) algorithm (Altschul et al., 1990). BLAST finds statistically significant similarities between sequences by evaluating alignments. Alignments are found using a variation of the dynamic programming approaches implemented in the Needleman-Wunsch (global alignment) and Smith-Waterman (local alignment) algorithms. The optimal alignment is calculated using statistics specifically developed for this (Korf et al.,

[1] `www.ncbi.nih.gov/Genbank/index.html`

2003), and reported as an E-value (the number of alignments expected by chance between a query sequence and all sequences in the database).

A lot of putative genes and unigenes have no sequence similarity to known genes, and their functions simply await discovery. In general however, the amount of new and inferred annotation is ever increasing. To facilitate the comparison of annotation between genes and organisms, and the communication between researchers, a group of model organism database providers has proposed, and is actively developing, a controlled vocabulary to use for gene annotation: the Gene Ontology, GO (Ashburner et al., 2000). The vocabulary is divided in three, to aid in the description of (1) the molecular functions of gene products, (2) their placement in and as cellular components, and (3) their participation in biological processes. Terms in each of the vocabularies are related to each other in a polyhierarchical manner, using 'is-a' and 'part-of' relationships. More formally, the terms and their relationships are structured as a directed acyclic graph. This means that a given term can relate to any other term, as long as this does not result in any term eventually being connected to itself. That is, as long as a series of terms and their relations does not produce a cycle. In practice, terms often relate to each in simple hierarchies where general terms divide into several distinct and more specific terms. Each GO term consists of a unique alphanumerical identifier, a common name, and a definition.

One can use such controlled vocabularies to test if any terms are overrepresented among groups of genes, compared to all genes measured. The basic idea of the testing is as follows: In an experiment, the expression of m genes are measured. Of these, m_c genes are tested as differentially regulated. Assume that out of the m genes, n of them have been annotated with the term "Protein biosynthesis". Among the group of differentially regulated genes, n_c of them are annotated with this term. The larger the ratio n_c/m_c compared to the ratio n/m, the less likely it is that this term appears among the differentially regulated genes simply due to random picking. Two approaches to calculate this likeliness are: Fisher's Exact test, which uses the hypergeometric distribution to calculate the probabilities of all outcomes when assuming random picking of terms, or the Chi-Square (goodness-of-fit) test, which uses the chi-square statistic to evaluate the difference between n_c/m_c and n/m (Upton and Cook, 2004).

4.2 Biochemical pathways and networks

To make the accumulating knowledge about the interactions between compounds in metabolic-, genetic information processing-, signal transduction-, and other pathways, easily accessible, databases/tools that store and order such information have been developed. Examples include the Kyoto Encyclopedia of Genes and Genomes, KEGG (Kanehisa et al., 2006), the Gene Map Annotator and Pathway Profiler, GenMAPP (Dahlquist et al., 2002), and MAPMAN (Thimm et al., 2004). For pathways stored in the databases, the integration of these with high throughput data is in general achieved by coloring each of the compounds in the pathway using a legend that relates to the expressions measured. For example, by coloring upregulated genes red, downregulated genes blue, non-changing genes grey, and genes not measured white. One does not necessarily need to measure the specific compounds of a pathway in order to get

valuable information. For example, if the genes coding for the specific enzymes of a metabolic pathway are identified, one can color the enzymes according to the regulation of these to get a representation of the transcriptional control of the metabolic pathway.

To store and represent the pathways in databases, more or less systematic languages are used to describe which compounds relate to each other and in what manner. A generalization of this is the Systems Biology Markup Language, SBML (Hucka et al., 2004). SBML is an XML-like computer-readable format developed for representing models of biochemical reaction networks. The generality of such a language allows pathways and models to be easily developed and shared by the community, and ancillary software modules to be quickly developed and integrated in the overall framework when needed. For example, having created a pathway one can simulate the dynamics of it by using the SBML differential equation solver, and this can then be compared to e.g. time-series measurements when they become available. The task of constructing a language that captures and concisely describes the intra- and extracellular networks of an organism at all levels (and between levels), is obviously not an easy one. For example, the flexibility of the XML approach (on which SBML is based) allows the same thing to be described in several ways which are not obviously identical (Wang et al., 2005). These authors therefore suggest that an approach that better captures the workings of the cell in an unique manner might be the reference-document format (RDF). In general, the development of language standards, ontologies, and databases that can store and analyze pathway information is one of the key efforts that can accelerate the overall rate of scientific discovery (Quackenbush, 2006) — and the optimal approach will ideally be selected based on what works and produces results and new insights.

An example of the additional insights gained when systematically storing networks in databases is the following: As shown by Guimerà and Amaral (2005), when exploring the overall topology of the entire network of interactions in a cell (as extracted from KEGG), functional modules appear — around 15 of them. That is, each network is divided into around 15 modules that only have a few connections between them. Furthermore, one can define compounds in these networks based on how they connect within and between modules and relate this to how they are affected by evolutionary constraints and pressures.

4.3 Research example: Functional classification of candidate genes and mapping to pathways (paper I)

Using the EST-sequence reads from the crop EST database CR-EST (Künne et al., 2000); genes on the 10k-array were annotated as follows. (1) By a BLASTX search against the non-redundant (nr) protein database at GenBank[2] to assign function by homology to similar genes, (2) by a TBLASTX search specifically against the Arabidopsis genome (The Arabidopsis Information Resource[3] to identify ortholog genes, and (3) by a BLASTN search against the

[2] `ftp.ncbi.nlm.nih.gov/blast/db`

[3] `ftp.arabidopsis.org/home/tair/home/tair/Sequences/blast_datasets/`

TIGR barley Gene Index database[4]. The TIGR barley Gene Index constitutes an assembly of all available barley ESTs into contigs, and TIGR annotates these contigs using GO terms and EC numbers (enzyme identification numbers) when possible. Hence, using the mapping obtained by blast between genes on the 10k-array and TIGR contigs, GO terms were assigned to 4868 genes on the array (47%), and EC numbers to 1446 genes on the array (14%). Besides the functional classification of genes based on GO terms, the candidate genes were also classified into 33 different functional groups resembling the major functional classes in the MAPMAN system (Thimm et al., 2004). For those candidate genes where GO terms were available, the two methods of functional classification were generally in agreement.

4.4 Research example: Overrepresented GO terms (paper I)

For each type of regulation of the candidate genes (each part of the Venn diagram in Fig. 3.2B in Sec. 3.3.6), Fisher's exact test was used to test for over-representation of GO-defined biological processes, cellular components, or molecular functions between these genes and the remaining genes on the array. A GO term was judged as significantly overrepresented when the Bonferroni-corrected p-value was below 0.05. For three types of regulation significantly over-represented GO terms were found: (1) Genes specifically upregulated in infected cells (7 GO terms), (2) genes upregulated in both infected and resistant cells (16 GO terms), and (3) genes downregulated in infected and resistant cells (5 GO terms). Focusing on the overrepresentation of 13 genes encoding protein components of the translation apparatus (GO term 0006412: protein biosynthesis) that are specifically upregulated in infected cells: Holcik and Sonenberg (2005) have linked sensing and response to stress to initiation of translation in terms of ribosome recruitment to internal ribosome-entry sites and cap-independent translation. Regulation of ribosome biogenesis is in yeast and mammalian cells controlled by the enzyme TOR (target of rapamycin) (Inoki et al., 2005). Signal cues from external or internal factors, including stress, modulate the activity of TOR which co-ordinately control events of cell growth, protein synthesis and cell death. We therefore speculate that increased translational activity in infected cells reflects aspects of balance between nutrient requirement and survival of the cell.

4.5 Research example: Biochemical pathways (paper I)

Among the 332 candidate genes identified as described in Sec. 3.3.6, only 64 of them have been assigned with an EC number. Since the following analysis is focused on identifying overall trends in the regulation of pathways, a larger number of false positives than the number we expect when FDR < 0.05 is

[4] `www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=barley`

acceptable. Therefore, for this analysis only, we set the significance threshold to FDR < 0.15, thus including 774 genes with 136 having an EC number. With this cutoff, metabolic pathways involving energy production and sugar conversions and phenylpropanoid biosynthesis appear to be regulated differently between infected, resistant and control cells. Below, the findings and hypotheses relating to mapping expression values to the starch and sucrose metabolic pathway (see Fig. 4.1) are discussed.

The enzyme UDP-glucose 6-dehydrogenase (EC 1.1.1.22; HH01A03) is upregulated in resistant cells. Presumably, this enzyme converts UDP-glucose to UDP-D-glucuronate for pectin biosynthesis. The coincident downregulation of poly-galacturonase (EC 3.2.1.15; HX04D02), involved in pectin degradation, suggests changes that could lead to cell wall fortification by pectin enrichment (Vogel et al., 2004). In line with this argument, Arabidopsis knock-outs of a pectate lyase gene (EC 4.2.2.2) have increased cell wall pectate content and show enhanced powdery mildew resistance (Vogel et al., 2002). Furthermore, a pectin esterase gene (HY10N07) is upregulated in resistant cells. This presumably also contributes to the cell wall strengthening by pectin demethylation (Vogel et al., 2004). The importance of pectin is further suggested by the fact that powdery mildew conidia contain pectinase presumably because this conveys an evolutionary advantage in pathogenesis, perhaps in facilitating cell wall penetration (Suzuki et al., 1998).

Central to powdery mildew nutrition is sucrose synthase (EC 2.4.1.13; HU03P12, HY10G10, and five others). We find that sucrose synthase is upregulated exclusively in infected cells (and sometimes even downregulated in resistant cells). Collinge et al. (1994) previously noted sucrose synthase upregulation following powdery mildew inoculation, but they used bulk leaf samples that also contained papilla-resistant cells; in these we found no such effect. Sucrose synthase catalyzes the reversible conversion of sucrose and UDP into fructose and UDP-glucose, and glucose is taken up preferentially by haustoria (Aked and Hall, 1993). This strongly suggests that sucrose synthase-mediated glucose generation would favor sugar uptake by the haustorium, which would, in turn, lead to carbon import from the underlying tissues to supply to the infected cell.

4.6 Text mining

The usual method for reporting new discoveries is through the publication of articles in reviewed journals. Most databases described here are initially based on what has been reported in articles — although as the databases become more and more known, results are also submitted directly. Considering that new results are produced at an ever increasing pace, and that the entire body of articles is already immense (PubMed, a life-science article search and retrieval system, covers around 12 million citations dating back to the mid-1960's[5]), the task of locating all relevant articles with respect to a given pathway or result is daunting. However, most of the articles are stored as text-files at the publishers web site, and consequently, with the right tools the whole body of literature could in principle be mined automatically for gene annotations, pathways, and relations

[5] `http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html`

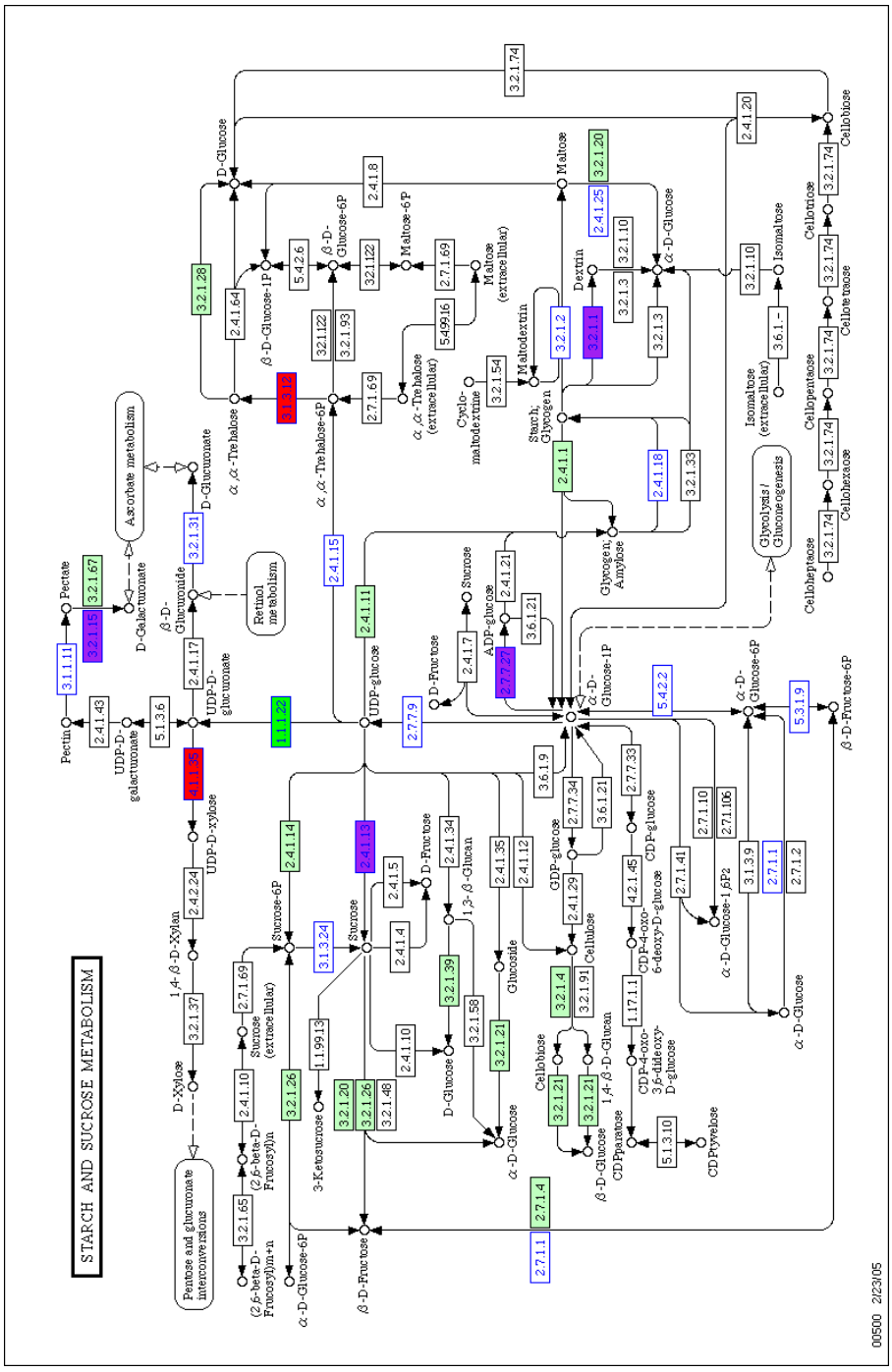

Figure 4.1: The starch and sucrose pathway as represented in KEGG. Enzymes represented in boxes with blue border are measured on the 10k-array, and those with black border are not. Boxes with white background have either not been found in barley (black border) or are not differentially regulated (blue border). For the differentially regulated genes, the following color codes are used for the background: dark green: upregulated in resistant cells, red: upregulated in infected cells, purple: upregulated in infected cells and either up- or downregulated in resistant cells.

between these and specific diseases, etc. A central issue in the development of such tools is the ability to recognize gene and protein names in article text. Researchers separated in time, research field, or organism studied have often devised different names for compounds that (later are shown to) have similar function. Renaming of such compounds is a slow process due to e.g. tradition. Therefore, there is a plethora of different names for the same gene or protein. In order to catalogue such names in a systematic manner, traditional classifiers such as Support Vector Machines and Hidden Markov Models have been used alone, or combined in an ensemble (Zhou et al., 2005). A simple use of such a classification is to search the entire body of literature for any mentioning of a group of genes of interest in relation to one or several keywords (Rubinstein and Simon, 2005). For example, which of the genes significantly regulated in barley epidermal cells infected with powdery mildew have previously been related to "barley", "powdery mildew", "infection", or a "stress" response in general? The long -term goal, however, is to develop tools that automatically extract also the relations between compounds (Zhou et al., 2005).

Conclusions and perspectives

Each of the methods to study single-cell gene expression, cell motility, antibody reactivity, and respiratory metabolism presented in this thesis has enabled the investigation of hitherto uncharted areas. The model systems they have been applied to generated interesting and novel data in their own right, and this guide further research in these areas. But they also function as proofs-of-concept, and application of the methods to other model systems are actively investigated. In addition, ways to integrate the methods presented here (and other methods), to improve the overall measurement coverage when studying systems of interest are also pursued.

For the single-cell transcript profiling of barley attacked by powdery mildew, follow-up experiments for several of the results have been designed. For example, having identified a group of protein synthesis genes that are specifically upregulated in infected cells, the immediate questions are: What controls this upregulation? Do several mechanisms that just converge at this point in the network, or is it just one factor, closely linked to the actions of the fungus. Our hypothesis concerning the actions of the gene encoding TOR and its involvement with these genes can only function as a guide for which independent experiments to perform next, or independent datasets to integrate this knowledge with. One approach to this would be to identify common motifs in the promoter region of the protein synthesis genes, to detect transcription factors that control the regulation. However, before embarking on this, it would be helpful to know whether these genes display a coordinated response not just at the 18 hai timepoint.

Because of this, and several other questions that were asked based on the findings in the dataset, we designed and conducted a time-course study in which we measured the gene expression in control and infected cells at 12, 15, 18, 21, 24, 36, and 48 hai. With these data, we did in fact find clear evidence that the protein synthesis genes display a complex, coordinated response, see Fig. 4.2. Since the barley genome has not been sequenced, the next difficulty is to determine the promoter regions for these genes. One experimental technique is to use a known sequence on the genome, e.g. an EST sequence, to design primers and then sequence the upstream region. Such a "genome walking" protocol was for example used by Mastyugin et al. (2004). However, it is very time consuming. Therefore, an initial, bioinformatic approach would be to take advantage of the fact that the rice genome, which is expected to be similar to the barley genome in many respects (Jaiswal et al., 2006), has been sequenced. One could therefore try to identify rice genes that are orthologues to the barley protein

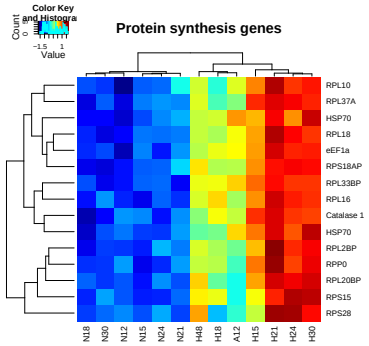

Figure 4.2: Expression profiles of protein synthesis genes.

synthesis genes, and then compare their promoter regions to detect common motifs. Doing this, and using a Gibbs-sampling approach (Aerts et al., 2005) for shared motif detection, the motif `GCGGCGGCG` was found in the rice promoter regions of 11 out of 13 of the protein synthesis genes. The next step, which has not yet been done, is to test whether this sequence motif is actually recognized by proteins (transcription factors) in infected cells. This could be done using an electrophoretic mobility shift assay. Also, this promoter is well described in mammalian systems (Egr-1 in humans and Igf-1 in rats). One could compare the cDNA sequences for the genes coding for these transcription factors to an assembly of barley unigenes, to see if any sequences with high similarity are present. If so, and if these barley unigenes also display the profile in Fig. 4.2 (even though that would be lucky since genes coding for transcription factors usually have very low expression), we might have found the control point. Another, and very simple experiment, regarding the observation that cell wall constituents seem to be upregulated in infected cells is to measure (using, e.g fourier-transform infrared absorption spectroscopy) the amount of pectin in infected cells and control cells and compare. Since pectin is a key structural component in cell walls this would confirm the transcriptome measurements at the metabolome level.

It would be interesting to apply the the single-cell transcript profiling method to cells walking on different surfaces. Besides having quantitative measures of the cells motility patterns, as calculated using the models developed in paper II, one would then also get measures of the basic regulatory state inside the cells. Preliminary gene expression studies for the response of motile cells to different surfaces have been performed by Klapperich and Bertozzi (2004). They show that genes involved in cell signaling, extracellular matrix remodeling, inflammation, angiogenesis, and hypoxia were all activated in cells on a collagen mesh compared to TCPS. In the case of developing methods for growing artificial skin, one could combine these methods, gathering knowledge on the optimal motility behavior of the cells needed for skin growth, and then use high-throughput gene expression measurements to help guide the engineering of cells that have these properties. There are of course ethical issues regarding such engineering.

The motility models developed was also a proof-of-principle. The basic point

of the paper was the development of a method for deriving motility models. Therefore, the next step is to apply the method to other cells to show the general applicability. This work has already begun for Dictyostelium discoideum (a soil-living amoeba) and fish scale keratocytes.

For the antibody microarray, research in several directions are well underway. Several datasets have been created and are currently analyzed, for example (1) non-obese diabetic mice have been immunized with HSP60, as protein in oil and as vector, to investigate disease mechanisms related to type 1 diabetes. And (2) antigen repertoires from patients with different subtypes of multiple sclerosis have been measured, and classifiers are being developed that can separate these subtypes. Furthermore, serum samples for a large study designed to investigate whether transplantation rejection responses can be detected before the transplantation procedure, are currently being collected. This again strikes at the theme concerning to what extent precise measurements of the state of a system before the perturbation (being in this case the immune system and the replacement of an organ) can be used to predict the response to the perturbation. Several transcript profiling studies on the response to transplantation as measured in peripheral blood have been performed (Jurcevic and Sacks, 2003) and such an approach could obviously be combined with antibody detection.

Finally, concerning the method to measure free NADH in vivo. There are two advantages to the use of a pulsed excitation regime: (1) Bleaching is reduced, and (2) it allows collection of time-gated data in addition to the steady-state spectra. Expanding the method initially introduced by Paul and Schneckenburger (1996), we made a series of measurements of NADH fluorescence lifetimes in different metabolic states. The measurements of fluorescence decays have an advantage over the steady-state measurements because the fluorescence lifetime of bound NADH is an order of magnitude longer than that of a free NADH. Consistent with this, our preliminary results indicated that the observed fluorescence of potato mitochondria has a longer lifetime in those states that have an increased proportion of bound NADH. Quantification of this qualitative result, however, requires a rigorous mathematical treatment and optimization of the experimental setup, and is under development.

Abbreviations

10k-array	barleyPGRC1 10K cDNA array
ADH	yeast alcohol dehydrogenase
ADP	adenosine diphosphate
ATP	adenosine triphosphate
BLAST	basic local alignment search tool
BSA	bovine serum albumin
CAD	cyclophosphamide accelerated diabetes
cDNA	complementary (or copy) DNA
CERN	European Particle Physics Laboratory
CV	coefficient of variation
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
dGTP	deoxyguanosine triphosphate
dTTP	deoxythymidine triphosphate
ddATP	dideoxyadenosine triphosphate
ddCTP	dideoxycytidine triphosphate
ddGTP	dideoxyguanosine triphosphate
ddTTP	dideoxythymidine triphosphate
DNA	deoxyribonucleic acid
EC	enzyme classification
ELISA	enzyme-linked immunosorbant assay
EST	expressed sequence tag
FDR	false discovery rate
FISH	fluorescence in situ hybridization
FN	false negative
FP	false positive
FWER	family wise error rate
GO	gene ontology
hai	hour after inoculation
HaCaT	human keratinocyte
HPLC	high-performance liquid chromatography

IgG	immunoglobulin subtype G
IgM	immunoglobulin subtype G
ITC	isothermal titration calorimetry
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCM	laser capture microdissection
LMPC	laser microdissection coupled to laser pressure catapulting
MDH	malate dehydrogenase, NAD^+-specific
mRNA	messenger RNA
MS	mass spectrometry
NAD	$NADH + NAD^+$
NAD^+	nicotinamide adenine dinucleotide (oxidised form)
NADH	nicotinamide adenine dinucleotide (reduced form)
NADP	$NADPH+NADP^+$
$NADP^+$	nicotinamide adenine dinucleotide phosphate (oxidised form)
NADPH	nicotinamide adenine dinucleotide phosphate (reduced form)
NAD(P)	NAD and NADP
ND_{in}	matrix facing rotenone-insensitive NADH dehydrogenase
NHDF	normal human dermal fibroblasts
OU	Ornstein-Uhlenbeck
PBS	phosphate buffered saline
PCR	polymerase chain reaction
RDF	reference-document format
RMSD	root mean square deviation
RNA	ribonucleic acid
SAGE	serial analysis of gene expression
SBML	systems biology markup language
SD	standard deviation
SDE	stochastic differential equation
SNP	single nucleotide polymorphism
TCPS	tissue culture grade polystyrene
TN	true negative
TP	true positive
XML	extensible markup language

References

Aebersold, R. and M. Mann (2003). Mass spectrometry-based proteomics. *Nature 422*, 198–207.

Aerts, S., P. Van Loo, G. Thijs, H. Mayer, Y. de Martin, R .and Moreau, and B. de Moor (2005). TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res. 33*, W393–W396.

Agius, S. C., A. G. Rasmusson, and I. M. Møller (2001). NAD(P) turnover in plant mitochondria. *Aust. J. Plant Physiol. 28*, 461–470.

Ahmed, A. A., M. Vias, N. G. Iyer, C. Caldas, and J. D. Brenton (2004). Microarray segmentation methods signifcantly influence data precision. *Nucleic Acids Res. 32*, e50.

Aked, J. and J. L. Hall (1993). The uptake of glucose, fructose and sucrose into pea powdery mildew (erysiphe pisi dc) from the apoplast of pea leaves. *New Phytol. 123*, 277–282.

Aloy, P. and R. B. Russel (2004). Ten thousand interactions for the moleculer biologist. *Nature Biotechnol. 22*, 1317–1321.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). Basic local alignment search tool. *J. Mol. Biol. 215*, 403–410.

Amaratunga, D. and J. Cabrera (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. John Wiley & Sons, Inc., New Jersey, USA.

Angulo, J. and J. Serra (2003). Automatic analysis of dna microarray images using mathematical morphology. *Bioinformatics 19*, 553–562.

Ashburner, M., C. A. Ball, J. A. Blake, D. B. D, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genet. 25*, 25–29.

Baldi, P. and A. Long (2001). A bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics 17*, 509–519.

Battogtokh, D., D. K. Asch, M. E. Case, J. Arnold, and H.-B. Schüttler (2002). An ensemble method for identifying regulatory circuits with special reference to the *qa* gene cluster of *Neurospora crassa*. *Proc. Natl. Acad. Sci. USA 99*, 16904–16909.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. 57*, 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist. 29*, 1165–1188.

Berg, H. C. (2000). Motile behavior of bacteria. *Phys. Today. 53*, 24–29.

Berg, H. C. (2003). *E. coli in Motion*. Springer-Verlag, New York.

Berg, H. C. and D. A. Brown (1972). Chemotaxis in escherichia coli analysed by three-dimensional tracking. *Nature 239*, 500–504.

Bevington, P. R. and D. K. Robinson (2004). *Data Reduction and Error Analysis for the physical sciences* (3rd ed.). McGraw-Hill, New York, USA.

Birnbaum, K., D. E. Shasha, J. Y. Wang, J. W. Jung, G. M. Lambert, D. W. Galbraith, and P. N. Benfey (2003). A gene expression map of the arabidopsis root. *Science 302*, 1956–1960.

Blatt, M., S. Wiseman, and E. Domany (1996). Superparamagnetic clustering of data. *Phys. Rev. Lett. 76*, 3251–3254.

Blinova, K., S. Carroll, S. Bose, A. V. Smirnov, J. J. Harvey, J. R. Knutson, and R. S. Balaban (2005). Distribution of mitochondrial NADH fluorescence lifetimes: steady-state kinetics of matrix NADH interactions. *Biochemistry 44*, 2585–2594.

Boguski, M. S., T. M. Lowe, and C. M. Tolstoshev (1993). dbEST–database for "expressed sequence tags". *Nature Genet. 4*, 332–333.

Bolstad, B. M., R. A. Irizarry, M. A. Å. 3, and T. P. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics 19*, 185–193.

Brandt, S., S. Kloska, T. Altmann, and J. Kehr (2002). Using array hybridization to monitor gene expression at the single cell level. *J. Exp. Bot. 53*, 2315–2323.

Brown, C. S., P. C. Goodwin, and P. K. Sorger (2001). Image metrics in the statistical analysis of dna microarray data. *Proc. Natl. Acad. Sci. USA 98*, 8944–8949.

Brown, K. S. and J. P. Sethna (2003). Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E 68*, 021904.

Bylesjö, M., D. Eriksson, A. Sjödin, M. Sjöstrm, S. Jansson, H. Antti, and J. Trygg (2005). MASQOT: a method for cDNA microarray spot quality control. *BMC Bioinformatics 6*, 250.

Caldo, R. A., D. Nettleton, and R. P. Wise (2004). Interaction-dependent gene expression in Mla-specified response to barley powdery mildew. *Plant Cell 16*, 2514–2528.

Celis, A., H. H. Rasmussen, P. Celis, B. Basse, J. B. Lauridsen, G. Ratz, B. Hein, M. Østergaard, H. Wolf, T. Ørntoft, and J. E. Celis (1999). Short-term culturing of low-grade superficial bladder transitional cell carcinomas leads to changes in the expression levels of several proteins involved in key cellular activities. *Electrophoresis 20*, 355–361.

Choe, S. E., M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon (2005). Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biology 6*, R16.

Ciotti, M. M. and N. O. Kaplan (1957). Procedures for determination of pyridine nucleotides. *Methods Enzymol. 3*, 890–899.

Close, T. J., S. I. Wanamaker, R. A. Caldo, S. M. Turner, D. A. Ashlock, J. A. Dickerson, W. R. A., G. J. Muehlbauer, A. Kleinhofs, and R. P. Wise (2004). A new resource for cereal genomics: 22K barley genechip comes of age. *Plant Physiol. 134*, 960–968.

Collinge, D. B., P. L. Gregersen, and H. T. Christensen (1994). *The Induction of Gene Expression in Response to Pathogenic Microbes*. Marcel Dekker, New York, USA.

Conway, J. P. and M. Kinter (2005). Proteomic and transcriptomic analyses of macrophages with an increased resistance to oxidized low density lipoprotein (oxLDL)-induced cytotoxicity generated by chronic exposure to oxLDL. *Mol. Cell. Proteomics 4*, 1522–1540.

Dahlquist, K. D., N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics 31*, 19–20.

Danø, S., P. G. Sørensen, and F. Hynne (1999). Sustained oscillations in living cells. *Nature 402*, 320–322.

de Lichtenberg, U., L. J. Jensen, S. Brunak, and P. Bork (2005). Dynamic complex formation during the yeast cell cycle. *Science 307*, 724–727.

Dodd, L. E., E. L. Korn, L. M. McShane, G. V. R. Chandramouli, and E. Y. Chuang (2004). Correcting log ratios for signal saturation in cdna microarrays. *Bioinformatics 20*, 2685–2693.

Dunn, M. J. and A. Gorg (2001). *Two-Dimensional Polyacrylamide Gel Electrophoresis for Proteome Analysis*, pp. 45–47. BIOS Scientific Publishers Limited, Oxford, UK.

Durbin, B. P., J. S. hardin, D. M. Hawkins, and D. M. Rocke (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics 18*, S105–S110.

Edwards, D. (2003). Non-linear normalization and background correction in one-channel cdna microarray studies. *Bioinformatics 19*, 825–833.

Ekstrøm, C. T., S. Bak, C. Kristensen, and M. Rudemo (2004). Spot shape modelling and data transformations for microarrays. *Bioinformatics 20*, 2270–2278.

Ellis, R. J. (2001). Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr. Opin. Struc. Biol. 11*, 114–119.

Elowitz, M. B., A. J. Levine, E. D. Siggia, and P. S. Swain (2002). Stochastic gene expression in a single cell. *Science 297*, 1183–1186.

Emmert-Buck, M. R., R. F. Bonner, P. D. Smith, R. F. Chuaqui, Z. Zhuang, S. R. Goldstein, R. A. Weiss, and L. A. Liotta (1996). Laser capture microdissection. *Science 274*, 998–1001.

Femino, A. M., F. S. Fay, K. Fogarty, and R. H. Singer (1998). Visualization of single RNA transcripts in situ. *Science 280*, 585–590.

Fields, S. and O.-k. Song (1989). A novel genetic system to detect proteinprotein interactions. *Nature 340*, 245–246.

Fodor, S. P., J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science 251*, 767–773.

Freund, R. J. and W. J. Wilson (2003). *Statistical Methods* (2nd ed.). Academic Press, San Diego, USA.

Fricke, W., P. S. Hinde, R. A. Leigh, and A. D. Tomos (1995). Vacuolar solutes in the upper epidermis of barley leaves. *Planta 196*, 40–49.

Gadegaard, N., S. Mosler, and N. B. Larsen (2003). Biomimetic polymer nanostructures by injection molding. *Macromol. Mater. Eng. 288*, 76–83.

Gail, M. H. and C. W. Boone (1970). The locomotion of mouse fibroblasts in tissue culture. *Biophys. J. 10*, 980–993.

Galinsky, V. L. (2003). Automatic registration of microarray images. i. rectangular grid. *Bioinformatics 19*, 1824–1831.

Gallagher, J. A., O. A. Koroleva, D. A. Tomos, J. F. Farrar, and C. J. Pollock (2001). Single cell analysis technique for comparison of specific mRNA abundance in plant cells. *J. Plant Physiol. 158*, 1089–1092.

Gaudet, C., W. A. Marganski, S. Kim, C. T. Brown, V. Gunderia, M. Dembo, and J. Y. Wong (2003). Influence of type i collagen surface density on fibroblast spreading, motility, and contractility. *Biophys. J. 85*, 3329–3335.

Gavin, A.-C., M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature 415*, 141–147.

Ge, H. (2000). UPA, a universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein-ligand interactions. *Nucleic Acids Res. 28*, e3.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol. 5*, R80.

Gerbal, F., P. Chaikin, Y. Rabin, and J. Prost (2000). An elastic analysis of listeria monocytogenes propulsion. *Biophys. J. 79*, 2259–2275.

Gjetting, T., T. L. Carver, L. Skøt, and M. F. Lyngkjær (2004). Differential gene expression in individual papilla-resistant and powdery mildew-infected barley epidermal cells. *Mol. Plant. Microbe. Interact. 17*, 729–737.

Guimerà, R. and L. A. N. Amaral (2005). Functional cartography of complex metabolic networks. *Nature 433*, 895–900.

Hagedorn, P. H., H. Flyvbjerg, and I. M. Møller (2004). Modelling NADH turnover in plant mitochondria. *Physiol. Plant. 120*, 370–385.

Hahn, A., J. Rahnenführer, P. Talwar, and T. Lengauer (2005). Confirmation of human protein interaction data by human expression data. *BMC Bioinformtics 6*, 112.

He, Y. D., H. Dai, E. E. Schadt, G. Cavet, S. W. Edwards, S. B. Stepaniants, S. Duenwald, R. Kleinhanz, A. R. Jones, D. D. Shoemaker, and R. B. Stoughton (2003). Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. *Bioinformatics 19*, 956–965.

Heid, C. A., J. Stevens, K. J. Livak, and P. M. Williams (1996). Real time quantitative pcr. *Genome Res. 6*, 986–994.

Higham, D. J. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review 43*, 525–546.

Holcik, M. and N. Sonenberg (2005). Translational control in stress and apoptosis. *Nature Rev. Mol. Cell Biol. 6*, 310–327.

Holloway, A. J., R. K. van Laar, R. W. Tothill, and D. D. L. Bowtell (2002). Options available—from start to finish—for obtaining data from DNA microarrays II. *Nature Genet. 32*(Suppl.), 481–489.

Horn, S., A. Lueking, D. Murphy, A. Staudt, C. Gutjahr, K. Schulte, A. König, M. Landsberger, H. Lehrach, S. B. Felix, and D. J. Cahill (2006). Profiling humoral autoimmune repertoire of dilated cardiomyopathy (dcm) patients and development of a disease-associated protein chip. *Proteomics 46*, 605–613.

Hu, S., D. A. Michels, M. A. Fazal, C. Ratisoontorn, M. L. Cunningham, and N. J. Dovichi (2004). Capillary sieving electrophoresis/micellar electrokinetic capillary chromatography for two-dimensional protein fingerprinting of single mammalian cells. *Anal. Chem. 76*, 4044–4049.

Huber, W., A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics 18*, S96–S104.

Hucka, M., A. Finney, B. Bornstein, S. M. Keating, B. E. Shapiro, J. Matthews, B. L. Kovitz, M. J. Schilstra, A. Funahashi, J. C. Doyle, and H. Kitano (2004). Evolving a lingua franca and associated software infrastructure for computational systems biology: the systems biology markup language (sbml) project. *Syst. Biol. 1*, 41–53.

Inoki, K., H. Ouyang, Y. Li, and K.-L. Guan (2005). Signaling by target of rapamycin proteins in cell growth control. *Microbiol. Mol. Biol. Rev. 69*, 79–100.

Iscove, N. N., M. Barbara, M. Gu, M. Gibson, C. Modi, and N. Winegarden (2002). Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nature Biotechnol. 20*, 940–943.

Jain, A. N., T. A. Tokuyasu, A. M. Snijders, R. Segraves, D. G. Albertson, and D. Pinkel (2002). Fully automatic quantification of microarray image data. *Genome Res. 12*, 325–332.

Jaiswal, P., J. Ni, I. Yap, D. Ware, W. Spooner, K. Youens-Clark, L. Ren, C. Liang, W. Zhao, K. Ratnapu, B. Faga, P. Canaran, M. Fogleman, C. Hebbard, S. Avraham, S. Schmidt, T. M. Casstevens, E. S. Buckler, L. Stein, and S. McCouch (2006). Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res. 34*, D717–D723.

Jewett, M. C., G. Hofmann, and J. Nielsen (2006). Fungal metabolite analysis in genomics and phenomics. *Curr. Opin. Biotechnol. 17*, 1–7.

Joanny, J. F., F. Jülicher, and J. Prost (2003). Motion of an adhesive gel in a swelling gradient: a mechanism for cell locomotion. *Phys. Rev. Lett. 90*, 168102.

Jung, H.-Y. and H.-G. Cho (2002). An automatic block and spot indexing with k-nearest neighbors graph for microarray image analysis. *Genome Res. 18*(Suppl. 2), S141–S151.

Jurcevic, S. and S. Sacks (2003). Gene microarrays in transplantation. *Curr. Opin. Nephrol. Hy. 12*(6), 577–579.

Kaempfer, R. (2003). RNA sensors: novel regulators of gene expression. *EMBO Rep. 4*, 1043–1047.

Kanehisa, M., S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res. 34*, D354–D357.

Karp, P. D. (2004). Call for an enzyme genomics initiative. *Genome Biology 5*, 401.

Karrer, E. A., J. E. Lincoln, S. Hogenhout, A. B. Bennett, R. M. Bostock, B. Martineau, W. J. Lucas, D. G. Gilchrist, and D. Alexander (1995). In situ isolation of mRNA from individual plant cells: Creation of cell-specific cDNA libraries. *Proc. Natl. Acad. Sci. USA 92*, 3814–3818.

Kidgell, C. and E. A. Winzeler (2005). Elucidating genetic diversity with oligonucleotide arrays. *Chromosome Res. 13*, 225–235.

Klapperich, C. M. and C. R. Bertozzi (2004). Global gene expression of cells attached to a tissue engineering scaffold. *Biomaterials 25*, 5631–5641.

Kononen, J., L. Bubendorf, A. Kallionimeni, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter, and O.-P. Kallionimeni (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Med. 4*, 844–847.

Korf, I., M. Yandell, and J. Bedell (2003). *BLAST*. O'Reilly & Associates, Sebastopol, USA.

Künne, C., M. Lange, T. Funke, H. Miehe, T. Thiel, I. Grosse, and U. Scholz (2000). CR-EST: a resource for crop ESTs. *Nucleic Acids Res. 33*, D619–D621.

Lai, Y., K. L. Feldman, and R. S. Clark (2005). Enzyme-linked immunosorbent assays (ELISAs). *Crit. Care Med. 33*, S433–S434.

Lauffenburger, D. A. and A. F. Horwitz (1996). Cell migration: a physically integrated molecular process. *Cell 84*, 359–369.

Leavitt, S. and E. Freire (2001). Direct measurement of protein binding energetics by isothermal titration calorimetry. *Curr. Opin. Struc. Biol. 11*, 560–566.

Levsky, J. M., S. M. Shenoy, R. C. Pezo, and R. H. Singer (2002). Single-cell gene expression profiling. *Science 297*, 836–840.

Levsky, J. M. and R. H. Singer (2003). Fluorescence in situ hybridization: Past, present and future. *J. Cell Sci. 116*, 2833–2838.

Liang, P. and A. B. Pardee (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science 257*, 967–971.

Liolios, K., N. Tavernarakis, P. Hugenholtz, , and N. C. Kyrpides (2006). The genomes on line database (GOLD) v.2: A monitor of genome projects worldwide. *Nucleic Acids Res. 34*, D332–D334.

Lopez, F., J. Rougemont, B. Loriod, A. Bourgeois, L. Loï, F. Bertucci, P. Hingamp, R. Houlgatte, and S. Granjeaud (2004). Feature extraction and signal processing for nylon dna microarrays. *BMC Genomics 5*, 38.

Machl, A. W., C. Schaab, and I. Ivanov (2002). Improving DNA array data quality by minimising 'neighbourhood' effects. *Nucleic Acids Res. 30*, e127.

Mann, M., R. C. Hendrickson, and A. Pandey (2001). Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem. 70*, 437–473.

Mastyugin, V., A. Mezentsev, W. X. Zhang, S. Ashkar, M. W. Dunn, and M. Laniado-Schwartzman (2004). Promoter activity and regulation of the corneal CYP4B1 gene by hypoxia. *J. Cell Biochem. 91*, 1218–1238.

Maxam, A. M. and W. Gilbert (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA 74*, 560–564.

Messer, R. (1994). *Linear Algebra: Gateway to Mathematics*. HarperCollins College Publishers, New York.

Orengo, C., D. Jones, and J. Thornton (Eds.) (2004). *Bioinformatics: Genes, Proteins and Computers*. BIOS Scientific Publishers Limited, Oxford, UK.

Ornstein, L. S. (1919). On the brownian motion. *Proc. Amst. 21*, 96–108.

Paul, R. and H. Schneckenburger (1996). Oxygen concentration and the oxidation-reduction state of yeast: Determination of free/ bound NADH and flavins by time-resolved spectroscopy. *Naturwissenschaften 83*, 32–35.

Penrose, R. (1956). On best approximate solution of linear matrix equations. *Proc. Camb. Philos. Soc. 52*, 17–19.

Ponti, A., M. Machacek, S. L. Gupton, C. M. Waterman-Storer, and G. Danuser (2004). Two distinct actin networks drive the protrusion of migrating cells. *Science 305*, 1782–1786.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1997). *Numerical Recipes in Fortran 90: The Art of Parallel Scientific Computing* (2nd ed.). Cambridge University Press.

Qian, H.-R. and S. Huang (2005). Comparison of false discovery rate methods in identifying genes with differential expression. *Genomics 86*, 495–503.

Quackenbush, J. (2006). Standardizing the standards. *Mol. Syst. Biol. doi:10.1038*, msb4100052.

Quintana, F. J., G. Getz, G. Hed, E. Domany, and I. R. Cohen (2003). Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bio-informatic approach to immune complexity. *J. Autoimmun. 21*, 65–75.

Rahnenführer, J. and D. Bozinov (2004). Hybrid clustering for microarray image analysis combining intensity and shape features. *BMC Bioinformatics 5*, 47.

Ridley, A. J., M. A. Schwartz, K. Burridge, R. A. Firtel, M. H. Ginsberg, G. Borisy, J. T. Parsons, and A. R. Horwitz (2003). Cell migration: integrating signals from front to back. *Science 302*, 1704–1709.

Robinson, W. H., C. Digennaro, W. Hueber, B. B. Haab, M. Kamachi, E. K. Dean, S. Fournel, D. Fong, M. C. Genovese, H. E. N. de Vegvar, K. Skriner, D. L. Hirschberg, R. I. Morris, S. Muller, G. J. Pruijn, W. J. V. Venrooij, J. S. Smolen, P. O. Brown, L. Steinman, and P. J. Utz (2002). Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nature Med. 8*, 295–301.

Romualdi, C., S. Trevisan, B. Celegato, G. Costa1, and G. Lanfranchi (2002). Improved detection of differentially expressed genes in microarray experiments through multiple scanning and image integration. *Nucleic Acids Res. 31*, e149.

Rubinstein, R. and I. Simon (2005). Milano custom annotation of microarray results using automatic literature searches. *BMC Bioinformatics 6*, 12.

Sanger, F., S. Nicklen, and A. R. Coulson (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA 74*, 5463–5467.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science 270*, 467–470.

Schütze, K. and G. Lahr (1998). Identification of expressed genes by laser-mediated manipulation of single cells. *Nature Biotechnol. 16*, 737–742.

Shaw, C. A. (2002). Theoretical consideration of amplification strategies. *Neurochem. Res. 27*(10), 1123–1131.

Shenderow, A. D. and M. P. Sheetz (1997). Inversely correlated cycles in speed and turning in ameba: on oscillatory model of cell locomotion. *Biophys. J. 72*, 2382–2389.

Shively, J. E. (2000). The chemistry of protein sequence analysis. *EXS. 88*, 99–117.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol. 3*, 3.

Smyth, G. K. (2005). *Limma: Linear Models for Microarray Data.* Springer, New York, USA.

Sterky, F. and J. Lundeberg (2000). Sequence analysis of genes and genomes. *J. Biotechnol. 76*, 1–31.

Stetter, M., G. Deco, and M. Dejori (2003). Large-scale computational modeling of genetic regulatory networks. *Artif. Intell. Rev. 10*, 75–93.

Stocchi, V., L. Cucchiarini, M. Magnani, L. Chiarantini, P. Palma, and G. Cresentini (1985). Simultaneous extraction and reversedphase high performance liquid chromatographic determination of adenine and pyridine nucleotides in human blood cells. *Anal. Biochem. 146*, 118–124.

Stolovitzky, G. and G. Cecchi (1996). Efficiency of DNA replication in the polymerase chain reaction. *Proc. Natl. Acad. Sci. USA 93*, 12947–12952.

Struglics, A., K. M. Fredlund, A. G. Rasmusson, and I. M. Møller (1993). The presence of a short redox chain in the membrane of potato tuber peroxisomes and the association of malate dehydrogenase with the membrane. *Physiol. Plant. 88*, 19–28.

Sumner, L. W., P. Mendes, and R. A. Dixon (2003). Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry 62*, 817–836.

Suzuki, S., Y. Komiya, T. Mitsui, S. Tsuyumu, H. Kunoh, T. L. W. Carver, and R. L. Nicholson (1998). Release of cell wall degrading enzymes from conidia of blumeria graminis on artificial substrata. *Ann. Phytopathol. Soc. Jpn. 64*, 160–167.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature 437*, 1299–1320.

Therneau, T., R. C. Tschumper, and D. Jelinek (2002). Sharpening spots: Correcting for bleedover in cDNA array images. *Mat. Biosci. 176*, 1–15.

Thimm, O., O. Bläsing, Y. Gibon, A. Nagel, S. Meyer, P. Krüger, J. Selbig, L. A. Müller, S. Y. Rhee, and M. Stitt (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J. 37*, 914–939.

Thygesen, H. H. and A. H. Zwinderman (2004). Comparing transformation methods for dna microarray data. *BMC Bioinformatics 5*, 77.

Tilstone, C. (2003). DNA microarrays: Vital statistics. *Nature 424*, 610–612.

Tohge, T., Y. Nishiyama, M. Y. Hirai, M. Yano, J.-i. Nakajima, M. Awazuhara, E. Inoue, H. Takahashi, D. B. Goodenowe, M. Kitayama, M. Noji, M. Yamazaki, and K. Saito (2005). Functional genomics by integrated analysis of metabolome and transcriptome of arabidopsis plants over-expressing an MYB transcription factor. *Plant J. 42*, 218–235.

Tomos, A. D. and R. A. Sharrock (2001). Cell sampling and analysis (SiCSA): metabolites measured at single cell resolution. *J. Exp. Bot. 52*, 623–630.

Uetz, P., L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg (2000). A comprehensive analysis of proteinprotein interactions in saccharomyces cerevisiae. *Nature 403*, 623–627.

Uhlenbeck, G. E. and L. S. Ornstein (1919). On the theory of brownian motion. *Phys. Rev. 36*, 823–841.

Upton, G. and I. Cook (2004). *Dictionary of Statistics*. Oxford University Press, New York, USA.

Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler (1995). Serial analysis of gene expression. *Science 270*, 484–487.

Vogel, J. P., T. K. Raab, C. Schiff, and S. C. Somerville (2002). PMR6, a pectate lyase-like gene required for powdery mildew susceptibility in arabidopsis. *Plant Cell 14*, 2095–2106.

Vogel, J. P., T. K. Raab, C. Somerville, and S. C. Somerville (2004). Mutations in PMR5 result in powdery mildew resistance and altered cell wall composition. *Plant J. 40*, 968–978.

Wakita, M., G. Nishimura, and M. Tamura (1995). Some characteristics of the fluorescence lifetime of reduced pyridine nucleotides in isolated mitochondria, isolated hepatocytes, and perfused rat liver in situ. *J. Biochem. 118*, 1151–1160.

Wang, D., S. Liu, B. J. Trummer, C. Deng, and A. Wang (2002). Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nature Biotechnol. 20*, 275–281.

Wang, X., S. Ghosh, and S.-W. Guo (2001). Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res. 29*, e75.

Wang, X., R. Gorlitsky, and J. S. Almeida (2005). From XML to RDF: how semantic web technologies will change the design of 'omic' standards. *Nature Biotechnol. 23*, 1099–1103.

Weckwerth, W., M. E. Loureiro, K. Wenzel, and O. Fiehn (2004). Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. USA 101*, 7809–7814.

Winter, P. C., G. I. Hickey, and H. L. Fletcher (2002). *Genetics* (second ed.). BIOS Scientific Publishers Limited, Oxford, UK.

Wit, E. and J. McClure (2003). Statistical adjustment of signal censoring in gene expression experiments. *Bioinformatics 19*, 1055–1060.

Workman, C., L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, H. B. Nielsen, H.-H. Saxild, C. Nielsen, S. Brunak, and S. Knudsen (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology 3*, 1–16.

Xia, Y., H. Yu, R. Jansen, M. Seringhaus, S. Baxter, D. Greenbaum, H. Zhao, and M. Gerstein (2004). Analyzing cellular biochemistry in terms of molecular networks. *Annu. Rev. Biochem. 73*, 1051–1087.

Yang, Y. H., J. M. Buckley, S. Dudoit, and T. P. Speed (2002). Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graph. Stat. 11*, 108–136.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed (2002). Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res. 30*, e15.

Zhou, G., D. Shen, J. S. Jie Zhang and, and S. Tan (2005). Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics 6*(Suppl. 1), S7.

Zhu, H. and M. Snyder (2003). Protein chip technology. *Curr. Opin. Chem. Biol. 7*, 55–63.

About RISØ

Mission

To promote an innovative and environmentally sustainable technological development within the areas of energy, industrial technology and bioproduction through research, innovation and advisory services.

Vision

RISØ's research shall extend the boundaries for the understanding of nature's processes and interactions right down to the molecular nanoscale. The results obtained shall set new trends for the development of sustainable technologies within the fields of energy, industrial technology and biotechnology. The efforts made shall benefit Danish society and lead to the development of new multi-billion industries.